一般発表 | Track 1: 自然言語処理・機械学習基礎

2026年3月2日(月) 15:30 ~ 17:40 | C会場

[9C] 説明可能AI・モデル解釈

座長:小口 正人(お茶の水女子大学)　コメンテータ:軽部 幸起(電気通信大学)

15:30 ~ 15:55

[9C-01] FeatUpによる高解像度特徴マップを用いたGrad-CAMの視認性向上

*門脇 健太郎[1]、仲山 旺雅[1]、亀谷 由隆[1] (1. 名城大学)

15:55 ~ 16:20

[9C-02] 追加学習前後のモデルの予測確率の変化に基づく情報の特異性推定

*疋田 知寛[1]、湯本 高行[1] (1. 兵庫県立大学)

16:20 ~ 16:45

[9C-03] 次元削減を用いたモデル非依存SHAP近似における計算効率と説明誤差のトレードオフ分析

*西条 啓佑[1]、宮森 恒[1] (1. 京都産業大学)

16:45 ~ 17:10

[9C-04] LLMのUnknown-Unknownを捉えるHuman-in-the-Loopエンティティマッチング

*岡山 紘汰[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学 情報学群、2. 筑波大学 図書館情報メディア系)

17:10 ~ 17:35

[9C-05] 複数エージェントの学習戦略を表現する解釈可能なルール集合の獲得

*今村 優志[1]、太田 学[2]、上野 史[2] (1. 岡山大学工学部工学科情報・電気・数理データサイエンス系、2. 岡山大学学術研究院環境生命自然科学学域)

FeatUpによる高解像度特徴マップを用いたGrad-CAMの視認性向上

門脇健太郎† 仲山 旺雅†† 亀谷 由隆††

† 名城大学大学院理工学研究科情報工学専攻 〒468–0073 愛知県名古屋市天白区塩釜口1丁目501
†† 名城大学情報工学部情報工学科 〒468–0073 愛知県名古屋市天白区塩釜口1丁目501
E-mail: ††ykameya@meijo-u.ac.jp

あらまし 深層学習モデルの判断根拠を可視化する手法として Grad-CAM [17] が広く利用されているが，生成されるヒートマップの解像度が低く，物体の詳細な境界や微小な特徴を捉えることが困難であるという課題がある．本研究では，モデル非依存のアップサンプリング技術である FeatUp [8] を Grad-CAM に統合し，高解像度な特徴マップに基づく可視化手法を提案する．提案手法では，学習済みモデルを固定したまま，ニューラルネットワークを用いて連続的な特徴空間を再構成し，高解像度なヒートマップを生成する．ImageNet および腎臓腫瘍 CT 画像を用いた実験により，提案手法が従来手法と比較して，忠実性を保ちながら視認性を向上させることを定性的・定量的に示す．

キーワード 説明可能 AI，深層学習，Grad-CAM，FeatUp

1 はじめに

近年，画像認識における機械学習の発展は著しく，Convolutional Neural Networks (以下，CNN) や Vision Transformer (以下，ViT) [6] を用いることで高精度な識別が可能となった．しかし，これらのモデルは複雑な非線形構造を持つブラックボックスであり，判断根拠の不透明さが医療や金融などの信頼性が求められる領域への応用の障壁となっている．この問題を解決するため，AI の判断根拠を可視化する説明可能な AI (XAI) [3] の研究が盛んに行われており，中でも勾配情報を用いて判断に寄与した領域を特定する Grad-CAM は，その汎用性の高さから広く普及している．

Grad-CAM は，モデルの最終畳み込み層の特徴マップを用いてヒートマップを生成する．しかし，深い層の特徴マップはストライド操作等により空間解像度が大幅に縮小（例：224 × 224 の入力に対し 7 × 7）されているため，拡大表示した際に結果がぼやけてしまい，対象物の詳細な構造を反映できないという課題がある．

本研究では，近年提案されたモデル非依存のアップサンプリングフレームワークである FeatUp を Grad-CAM に統合することで，この低解像度問題を解決し，高精度かつ視認性の高い可視化を実現することを目的とする．本研究の主な貢献は以下の通りである．

- 高解像度 Grad-CAM の提案: FeatUp の Implicit 表現を活用し，既存モデルの再学習なしに高精細なヒートマップを生成する手法を提案した．
- 多角的な評価: CNN および ViT ベースの複数モデルに対し，ImageNet を用いた一般画像評価と，KiTS23 データセットを用いた腎臓腫瘍 CT 画像への適用評価を行い，提案手法の有効性を検証した．
- 定量的精度の向上: AOPC [7] や Insertion スコア [16] による忠実性評価，IoU および BF スコア [2] を用いた領域特定精度の評価で従来手法を上回る性能を確認した．

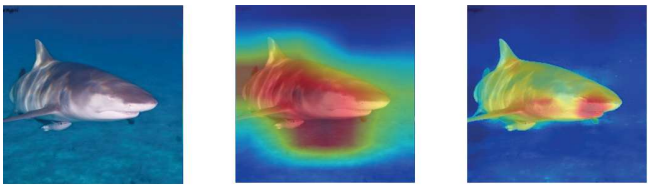

図 1 Grad-CAM と提案手法（Grad-CAM + FeatUp）のヒートマップの比較

本論文の構成は，まず第 1 章で研究背景について述べる [1]．第 2 章で本研究の基礎となる Grad-CAM および FeatUp の概要について述べる．第 3 章では，提案手法の詳細を定式化とともに説明する．第 4 章では，実験設定，使用データセット，および評価指標について述べる．第 5 章では，ImageNet に対する実験結果とその考察を示し，続く第 6 章では，医療ドメインへの応用として腎臓 CT 画像を用いた実験結果について論じる．最後に第 7 章で本研究の結論と今後の課題をまとめる．

2 準備および関連技術

2.1 Grad-CAM

Grad-CAM は，特定のクラス c に対する最終層 A^k の勾配を重み α_k^c として用い，以下の線形結合によってヒートマップ L^c を生成する：

1：本論文で提案している手法は The 2025 Principle and Practice of Data and Knowledge Acquisition Workshop (PKAW 2025) で最初に提案したものである [12]．本論文では PKAW 2025 発表論文に対して，FeatUp の詳細に関する記述，ImageNet データセットに対するセグメンテーション系指標での評価実験の結果と考察，腎臓 CT 画像データセットに対する評価実験の結果と考察を加えている．

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{1}$$

ここで，$L^c_{\text{Grad-CAM}}$ は空間解像度が低いため，最終的な出力にはバイリニア補間等の単純な拡大が必要となる．

2.2 FeatUp

FeatUp は，深層学習モデルの特徴マップをアップサンプリングするためのモデル非依存なフレームワークである．FeatUp には Joint Bilateral Upsampling (JBU) ベースの手法と Implicit ベースの手法の 2 種類が存在するが，本研究ではより柔軟な表現が可能である Implicit にのみ着目する．Implicit FeatUp (以下，単に FeatUp) は，学習済みモデル（バックボーン）のパラメータを固定したまま，特徴マップを再構成するための小さな MLP を学習する．この MLP は，画像座標と元の画像の特徴（RGB 値など）を入力とし，高解像度な特徴ベクトルを出力する連続関数として機能する．学習時には，入力画像に様々なデータ拡張（ジッター）を施し，それらに対するバックボーンの出力と，MLP による再構成特徴量との整合性を取ることで，高解像度化特徴マップを取得する．より詳細な説明は 5 章にて行う．

3 提案手法

本章では，Grad-CAM における特徴マップの低解像度問題を解決するために，特徴マップ高解像度化手法である FeatUp を統合した可視化手法を提案する [2]．

3.1 Implicit ニューラルネットワークによる特徴の定式化

Neural Radiance Fields (NeRF) [14] の思想に着想を得て，FeatUp は特徴マップを，正規化された 2 次元空間座標上で定義された連続関数として表現する．入力画像上の座標を $p = (x, y) \in [-1, 1]^2$ とし，その位置における RGB のカラーベクトルを c とする．このとき，座標 p における予測特徴ベクトル $\hat{f}(p)$ は，MLP を用いて次のように予測される．

$$\hat{f}(p) = \text{MLP}(\gamma(p) : \gamma(c) : c) \tag{2}$$

ここで，: はチャネル方向の結合を表す．また，$\gamma(\cdot)$ はフーリエエンコーディング関数であり，入力ベクトル z に対して以下のように定義される．

$$\gamma(z) = [\sin(\hat{\omega}_1 z), \cos(\hat{\omega}_1 z), \ldots, \sin(\hat{\omega}_L z), \cos(\hat{\omega}_L z)] \tag{3}$$

ここで，$\hat{\omega} = \{\hat{\omega}_1, \ldots, \hat{\omega}_L\}$ はエンコーディングに使用される周波数成分の集合である．本実験では L は 15 と設定している．フーリエエンコーディングは低次元の座標入力を高周波成分を含む空間へ写像するため，MLP の高周波成分（画像で言うところの境界線・エッジ）に関する表現力が向上する．この MLP は，フーリエエンコードされた空間座標，フーリエエンコードされた色特徴，および画素値を入力として受け取る．これにより，画像の特徴を考慮しつつ，解像度に依存しない形式で高解像度な特徴を推定することが可能となる．

3.2 多視点一貫性に基づく学習と損失関数設計

FeatUp では，単一の入力画像 I に対して，微小な変換（拡大，平行移動など）を施した複数の入力 $I_t = \mathcal{T}_t(I)$ $(t \in \mathcal{T})$ を用いて学習を行う．ここで，$\mathcal{T}$ は画像に適用されるランダムな変換の集合を表し，$\mathcal{T}_t$ は各変換演算子を指す．このとき，各変換画像に対するバックボーンネットワークの出力特徴を

$$F_t = \Phi(\mathcal{T}_t(I)) \tag{4}$$

とする．一方，FeatUp により得られる高解像度特徴マップを F_{hr} とし，これに同一の変換 $\mathcal{T}_t$ を適用した後，バックボーンのダウンサンプリング操作 $\sigma_\downarrow(\cdot)$ を施すことで，低解像度特徴 $\hat{F}_t$ を得る．

$$\hat{F}_t = \sigma_\downarrow(\mathcal{T}_t(F_{\text{hr}})) \tag{5}$$

3.2.1 再構成損失

FeatUp では，単純な平均二乗誤差（MSE）ではなく，各空間位置における再構成の信頼度を動的に推定するヘテロスケダスティック回帰に基づく再構成損失（Reconstruction Loss）を用いる．具体的には，MLP によって各画素位置ごとにスカラーのスケール s を推定し，以下の損失関数を最小化する．

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{t\in\mathcal{T},\, i,j}\left[\frac{1}{2s^2_{i,j}}\left\|F_t(i,j) - \hat{F}_t(i,j)\right\|_2^2 + \log(s_{i,j})\right] \tag{6}$$

ここで，(i, j) は高解像度特徴マップ F_{hr} における水平方向および垂直方向の画素位置を表す．$s_{i,j}$ は空間位置 (i, j) における学習可能なスケールであり，再構成誤差に対する重み付けとして機能する．この形式は，ノイズや外れ値を含む領域の影響を抑制しつつ，信頼性の高い領域を優先的に学習する効果を持つ．$\log(s)$ 項は，スケールが過度に大きくなることを防ぐ正則化として機能する．

3.2.2 Magnitude Loss

本手法は高周波成分を過剰に生成する傾向があるため，特徴量の発散を防ぐ目的で，高解像度特徴のノルムを教師特徴の大きさに一致させる Magnitude Loss を導入する．

$$\mathcal{L}_{\text{mag}} = \mathbb{E}_{i,j}\left[\left(\|F_{\text{hr}}(i,j)\| - \|F_{\text{target}}(i,j)\|\right)^2\right] \tag{7}$$

これにより，高解像度化による数値的不安定性を抑え，意味的に整合した特徴スケールを維持する．

3.2.3 Total Variation 正則化

高解像度特徴のノルムに対して局所的な滑らかさを課すため，特徴量の大きさ（magnitude）に対して Total Variation 正則化を導入する．

$$\begin{aligned}\mathcal{L}_{\text{tv}} = &\sum_{i,j}\left(\|F_{\text{hr}}(i,j)\| - \|F_{\text{hr}}(i-1,j)\|\right)^2 \\ &+ \sum_{i,j}\left(\|F_{\text{hr}}(i,j)\| - \|F_{\text{hr}}(i,j-1)\|\right)^2\end{aligned} \tag{8}$$

2：本節では FeatUp を説明する数式がたびたび登場するが，これは FeatUp の実装（https://github.com/mhamilton723/FeatUp）のソースコードを著者が数式化したものである．

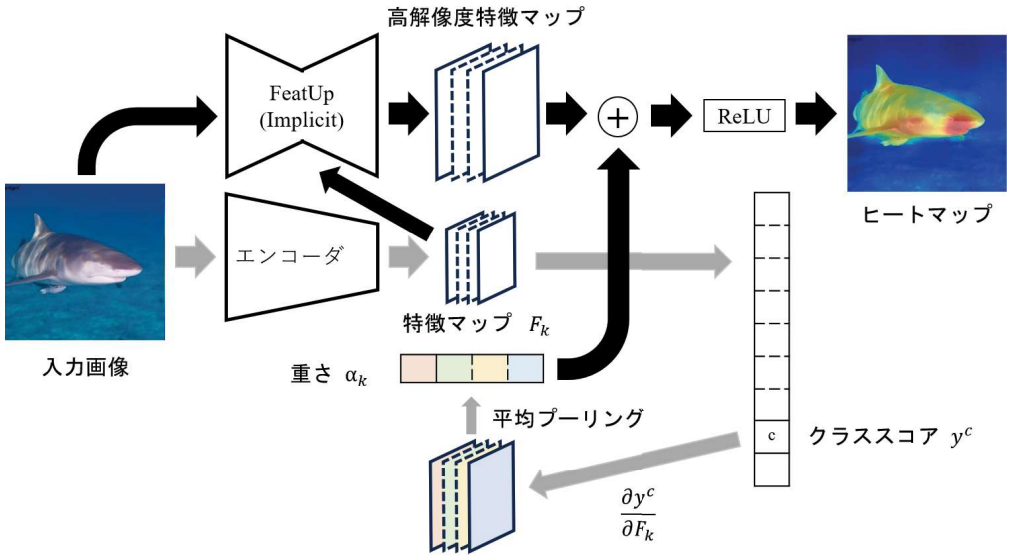

図 2　提案手法の概要図

この正則化により，空間的に不連続なスパイク状の応答が抑制され，視覚的に滑らかな高解像度特徴マップが得られる．

3.2.4　Blur Pin Loss

さらに，高解像度特徴が意味的構造を保ったまま過度な高周波成分を含まないよう，ガウシアンブラーを用いた Blur Pin Loss を導入する．

$$\mathcal{L}_{\text{blur}} = \mathbb{E}\left[\|\text{GaussianBlur}(F_{\text{hr}}) - F_{\text{hr}}\|_2^2\right] \tag{9}$$

この損失により，細部構造を保持しつつ，ノイズ的な成分の抑制が可能となる．

3.2.5　最終的な損失関数

以上の損失を重み付き和として統合し，FeatUp の最終的な学習目的関数を次式で定義する．

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} + \lambda_{\text{blur}}\mathcal{L}_{\text{blur}} \tag{10}$$

これにより，空間的・意味的一貫性を満たす高解像度特徴が学習される．

3.3　FeatUp と Grad-CAM の統合

次に，FeatUp によって得られた高解像度特徴表現を Grad-CAM に統合する手法について述べる．本研究で提案する統合手法は，Grad-CAM が本来持つクラス識別能力や説明性を維持したまま，可視化結果の空間解像度と視認性を向上させることを目的とする．そのため，Grad-CAM の計算過程そのものを変更するのではなく，Grad-CAM が参照する特徴マップを FeatUp によって高解像度化するという設計を採用する．2 に提案手法の概要図を示す．

3.3.1　勾配重みの算出

対象クラス c のスコアを y^c とすると，Grad-CAM のチャネル重み α_k^c は以下で与えられる．

$$\alpha_k^c = \frac{1}{Z}\sum_{i,j}\frac{\partial y^c}{\partial F_k(i,j)} \tag{11}$$

この重みはチャネル方向の重要度を表すスカラー量であり，空間解像度そのものとは無関係である．そのため，本手法では α_k^c を従来の Grad-CAM と同様に低解像度特徴マップから算出する．

3.3.2　高解像度ヒートマップの生成

次に，FeatUp によって生成された高解像度特徴マップを用いて，Grad-CAM のヒートマップを高解像度で構成する．FeatUp により得られる特徴は，連続座標 p において定義された特徴関数 $\hat{f}_k(p)$ として表される．

これらの高解像度特徴に対して，前節で算出したチャネル重み α_k^c を適用し，次式により高解像度 Grad-CAM を定義する．

$$L_{\text{hr}}^c(p) = \text{ReLU}\left(\sum_k \alpha_k^c \hat{f}_k(p)\right) \tag{12}$$

この定式化は，Grad-CAM が持つ，クラス判定に寄与する領域を特定する能力を保持したまま，得られるヒートマップの解像度を向上させる．特に，FeatUp によって復元された高解像度特徴は，単なる補間ではなく，バックボーンモデルの出力と整合性を保つように学習されているため，意味的な欠損が少ない補間となる．

4　実験方法

本章では，提案手法の有効性を検証するための実験環境，データセット，および評価手順について述べる．

4.1　データセット

汎用的な物体認識タスクとして ImageNet，ドメイン固有の実応用として腎腫瘍 CT 画像セット KiTS23 を用いる．

4.1.1　ImageNet および ImageNet-S

ImageNet (ILSVRC 2012) のバリデーションデータから各クラス 5 枚，計 5,000 枚を使用する．また，ピクセル単位の評価を可能にするため，セグメンテーションアノテーションが付与された ImageNet-S [9] を導入し，ヒートマップと正解領域

表 1　学習および正則化パラメータの設定

学習パラメータ		FeatUp 正則化項の重み		
項目	設定値	Loss	CNN 系	ViT 系
最適化器	Adam/SGD	Magnitude	0.05	0.05
学習率	1×10^{-4}	TV	0.001	0.002
エポック数	50	Blur	0.5	0.1

表 2　各モデルの識別性能（KiTS23 データセット）

モデル	F1	AUROC	AUPRC
VGG19	0.629	0.840	0.654
ResNet50	0.716	0.853	0.699
ViT-L/16	0.686	0.881	0.711
Swin-L	0.732	0.888	0.763

の空間的一致度を算出する．

4.1.2　腎腫瘍データセット：KiTS23

KiTS23 [11] より，腎臓が含まれる 2D スライスを抽出し，全サンプルを右腎に統一した．同一患者が複数のセットに混在しないよう，患者単位で 6：1：1 に分割した．正解領域（マスク）には，研修医が作成し専門医が監修した公式アノテーションを使用する．汎化性能向上のため，田中ら [19] と同様に，平行移動・回転・せん断・スケーリングを含む 5 種類の画像拡張を行い，データを 15 倍に拡張した．

4.2　バックボーンモデルと学習設定

バックボーンとして VGG19 [18]，ResNet50 [10]，ViT-L，Swin Transformer-L（以下，Swin）[13] の 4 種を用いる．ImageNet は事前学習済み重みを用い，KiTS23 は表 1 の設定でファインチューニングを行った．各モデルの識別性能を表 2 に示す．AUROC は ROC 曲線下の面積，AUPRC は Precision-Recall 曲線下の面積を指している．

4.3　FeatUp および評価手順

4.3.1　FeatUp の実行設定

各テスト画像に対し，NAdam（学習率=1×10^{-3}）を用いて 2,000 ステップの反復計算を行う．1 枚につき 20 個のジッター画像を生成し，表 1 右欄の重み設定で損失関数を最適化する．

4.3.2　評 価 指 標

生成されたヒートマップを以下の指標で評価する．

- **忠実性**: AOPC および Insertion スコアを用い，モデルの判断根拠を正しく反映しているか評価する．
  - **AOPC**: モデルの予測に対して影響度が大きい画素を順に削除（摂動）した際の，予測スコアの減少量を測定する指標である．ステップ L までの平均減少量として次式で定義され，この値が大きいほど，ヒートマップが重要な領域を正確に特定できていることを示す：

$$\mathrm{AOPC} = \frac{1}{L+1} \sum_{k=0}^{L} f(\mathbf{x}^{(0)}) - f(\mathbf{x}^{(k)}) \qquad (13)$$

  - **Insertion スコア**: 全ての画素値が 0 である一様な画像から，影響度が高いと推定された画素を順に挿入していった際の予測スコアの推移を評価する指標である．予測スコア曲線の面積として計算され，この値が高いほど，少ない画素数でモデルの予測を再現できていることを意味する．
- **視認性・特定精度**: 生成されたヒートマップに対し，Otsu の二値化法 [15] により背景を分離する．その後，正解マスクとの IoU および BF スコアを算出し，領域特定精度を評価する．
  - **IoU**: 提案手法が抽出した注目領域 R_{det} と，正解領域 R_{gt} の重なり度合いを評価する指標で，次式で定義される：

$$\mathrm{IoU} = \frac{|R_{det} \cap R_{gt}|}{|R_{det} \cup R_{gt}|} \qquad (14)$$

  - **BF スコア**: 抽出された領域の境界線の正確さを評価する指標である．正解境界と予測境界の距離に基づく Precision と Recall の調和平均として定義され，境界線をいかに考慮した可視化が取得できているかを定量化する：

$$\text{BF スコア} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (15)$$

また，定性評価を併せて行い，輪郭への適合性を確認する．

5　ImageNet に対する実験結果と考察

本章では，提案手法（FeatUp + Grad-CAM）の有効性を ImageNet データセットを用いて定量・定性の両面から検証する．

5.1　定性評価と視覚的考察

図 3 に，各アーキテクチャにおける従来の Grad-CAM と提案手法の比較を示す．

従来の Grad-CAM では最終畳み込み層の低解像度に起因し，物体の境界周辺に広範なにじみが発生していた．これにより，モデルが物体のどの範囲を識別根拠としたかの厳密な判断が困難であった．これに対し提案手法では，物体の輪郭に沿って活性値がまとまった様子が確認された．提案手法によるヒートマップの特徴として，対象物体の特徴的な点（例えば，鳥の頭部や尾羽など）を捉える傾向がある．また，複数の物体が近接して存在するシーンにおいて，従来手法ではそれらを一つの連続した活性領域として統合してしまう課題があったが，提案手法では個々の物体の境界を分離して抽出する能力が向上していることが示された．

ViT および Swin においては，パッチ単位の処理に起因する空間的な不連続性やノイズが課題であったが，提案手法によりこれらのアーティファクトが抑制され，空間的に滑らかなヒートマップが得られた．

5.2　摂動系指標による忠実性評価

表 3 に，モデル予測スコアの変化に基づく忠実性（Fidelity）の評価結果を示す．

AOPC に関しては，CNN 系モデル（VGG19, ResNet50）および Swin において，提案手法がベースラインを上回った．特に ResNet50 の高確信度グループ（Conf. $>$ 90%）では 0.828

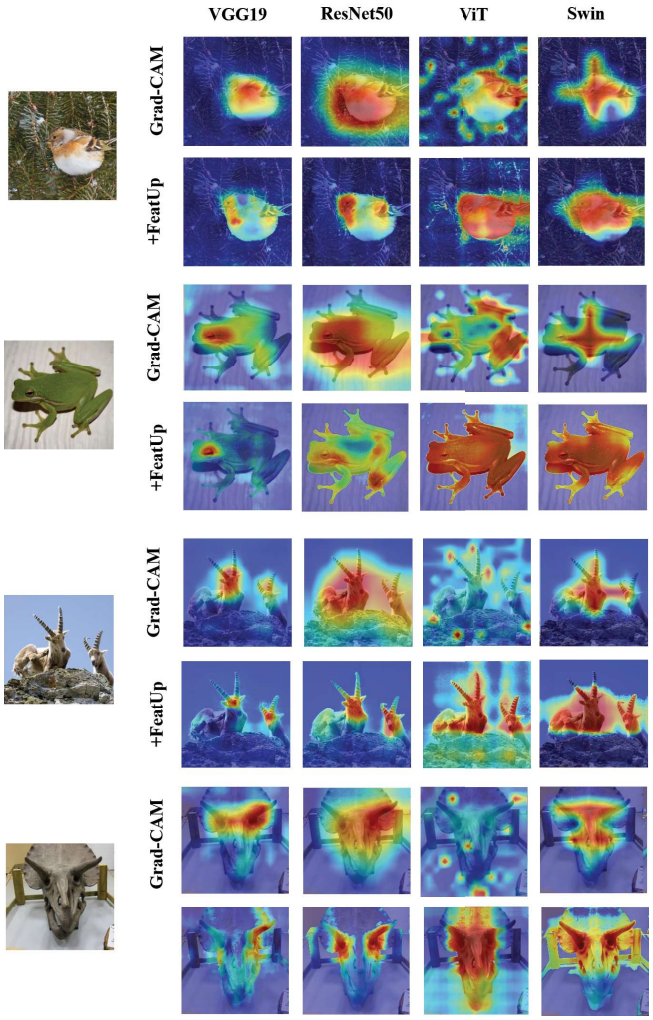

図 3　ImageNet におけるヒートマップの例

から 0.853 へと向上しており，重要な画素を的確に保持できている．

一方で，CNN ベースでは Insertion スコアが僅かに低下した．これはヒートマップがシャープになりすぎたことで，予測に必要な面積をカバーするまでのステップ数が増加したためと推測される．対照的に，ViT-L では Insertion スコアが向上（0.547 から 0.611）した．これは，Transformer において，FeatUp の高解像度化が決定的な活性領域をピクセルレベルで特定できていることを示唆している．

5.3　セグメンテーション指標による評価

図 4 に，ImageNet の検証データセットにおける IoU および BF スコアの平均値を示す．IoU に関しては，全てのバックボーンモデルにおいて提案手法がベースラインを上回った．向上幅に注目すると，ViT-L では 0.332 から 0.426（約 28%増），Swin-L では 0.371 から 0.463（約 25%増）となっており，Transformer ベースのモデルにおいて特に改善が確認された．

さらに，境界の再現性を評価する BF スコア においては，IoU を上回る性能向上が確認された．ResNet50 では 0.073 から 0.264 へと約 3.6 倍に向上し，ViT-L および Swin-L においてもベースラインの約 2 倍近いスコアを記録した．この結果

表 3　ImageNet におけるヒートマップの忠実性評価

モデル	手法	AOPC ↑	Insertion ↑
VGG19	Grad-CAM (All)	0.584	**0.398**
	+ FeatUp (All)	**0.586**	0.357
	Grad-CAM (Conf. > 90%)	0.856	**0.644**
	+ FeatUp (Conf. > 90%)	**0.859**	0.566
ResNet50	Grad-CAM (All)	0.615	**0.480**
	+ FeatUp (All)	**0.629**	0.459
	Grad-CAM (Conf. > 90%)	0.828	**0.712**
	+ FeatUp (Conf. > 90%)	**0.853**	0.688
ViT-L	Grad-CAM (All)	**0.559**	0.547
	+ FeatUp (All)	0.516	**0.611**
	Grad-CAM (Conf. > 90%)	**0.671**	0.670
	+ FeatUp (Conf. > 90%)	0.618	**0.749**
Swin-L	Grad-CAM (All)	0.452	**0.536**
	+ FeatUp (All)	**0.479**	0.519
	Grad-CAM (Conf. > 90%)	0.529	**0.683**
	+ FeatUp (Conf. > 90%)	**0.563**	0.659

表 4　ImageNet におけるセグメンテーション指標評価

モデル	手法	IoU ↑	BF-socre ↑
VGG19	Grad-CAM	0.185	0.070
	+ FeatUp	**0.212**	**0.124**
ResNet50	Grad-CAM	0.410	0.073
	+ FeatUp	**0.447**	**0.264**
ViT-L	Grad-CAM	0.332	0.115
	+ FeatUp	**0.426**	**0.211**
Swin-L	Grad-CAM	0.371	0.099
	+ FeatUp	**0.463**	**0.217**

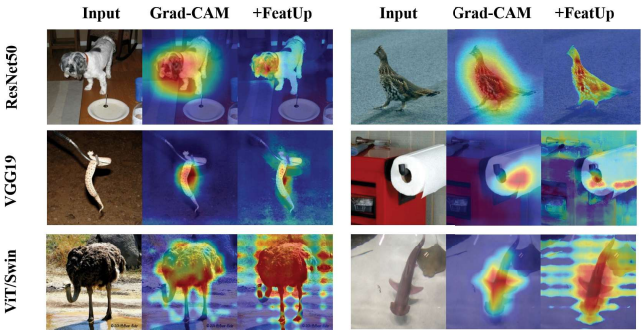

図 4　品質が良くないヒートマップの例

は，提案手法が物体のおおまかな位置を特定するだけでなく，物体の輪郭を精密に捉えていることを定量的に裏付けている．

5.4　ImageNet における課題事例の分析

図 5.4 に課題事例を示す．提案手法は多くの事例で視認性を向上させたが，一部で背景への過剰反応や活性値の均一化といった課題も確認された．低確信度サンプルの幾何変換に対する不安定さが主な要因と推察される．また，Transformer 系モデルではパッチ間の不連続性に起因する格子状ノイズ等のアーティファクトが発生しており，アーキテクチャに応じた正則化の検討が今後の課題である．

6 腎臓 CT 画像に対する実験評価と考察

医療ドメインの実用性を検証するため，KiTS23 データセットを用いた評価を行った．

6.1 定性評価と局在化の特性

図 6.1 に，各モデルにおける従来の Grad-CAM と提案手法（+ FeatUp）の比較を示す．ここで，Ground Truth の緑は腎臓の位置を示し，赤は腫瘍の位置を示している．全体的な傾向として，従来の Grad-CAM ではバックボーンの解像度不足に起因し，活性領域が腎臓周辺の脂肪組織や隣接臓器へとはみ出すことが多かったが，提案手法では臓器の構造に適合した，シャープな可視化が可能となった．

各モデルの詳細な挙動を以下に述べる．

- **VGG19**: 予測の正誤に関わらず，従来の時点で腎臓領域を正確に捉えられておらず，提案手法においてもその傾向が継承された．画像全体が不自然にハイライトされる不安定な挙動は，VGG19 の受容野や特徴抽出能力が，微細な医療画像の構造を捉えるには不十分であり，FeatUp の復元元となる特徴マップ自体に有用な情報が欠落していたことを示唆している．
- **ResNet50**: 正解例（TP, TN）において，最も劇的な改善が見られた．従来の Grad-CAM では腫瘍周辺に広範に分布していた活性が，提案手法により腫瘍部位のみに厳密に限定された．これは診断支援における視認性を大きく向上させる結果である．一方で不正解例では，誤った注目領域を鮮明に強調する結果となり，モデルの誤判断を忠実に可視化していると言える．
- **ViT / Swin-L**: 正解例において良好な局在化を示した．特に ViT では，従来の Grad-CAM が腫瘍中心から僅かに解離していた位置を，正確に腫瘍位置へ補正する様子が確認された．これは，FeatUp が単なるエッジ強調ではなく，MLP を通じた学習によりセマンティックな情報を再構成している証左である．ただし，Swin-L では ViT と比較してヒートマップに僅かな滲みが生じる傾向があり，これは ImageNet での実験結果とも整合する．

6.2 摂動系指標による忠実性評価

表 5 に摂動系指標の結果を示す．AOPC に関しては，ResNet50 においてベースラインの 0.172 から 0.271 へと大幅な向上が確認された．一方で，VGG19，ViT-L，Swin-L においてはベースラインを下回る結果となった．Insertion スコアについては，VGG19 において 0.482 から 0.541 への向上が見られた．その他のモデル（ResNet50, ViT-L, Swin-L）においては，ベースラインと比較してスコアが僅かに減少，あるいは同等の数値を記録した．特に Transformer ベースのモデル（ViT-L, Swin-L）では，手法を問わず 0.9 前後の高い Insertion スコアを維持していることが確認された．

ImageNet では Insertion スコアがモデルや手法によって大きく変動したが，腎臓 CT においては Transformer ベースのモデルで 0.9 前後の高い値で飽和する傾向が見られた．これは，一般画像が多様な背景ノイズを含むのに対し，CT 画像は背景の大部分が低輝度（黒色）であり，モデルが少数の重要な画素

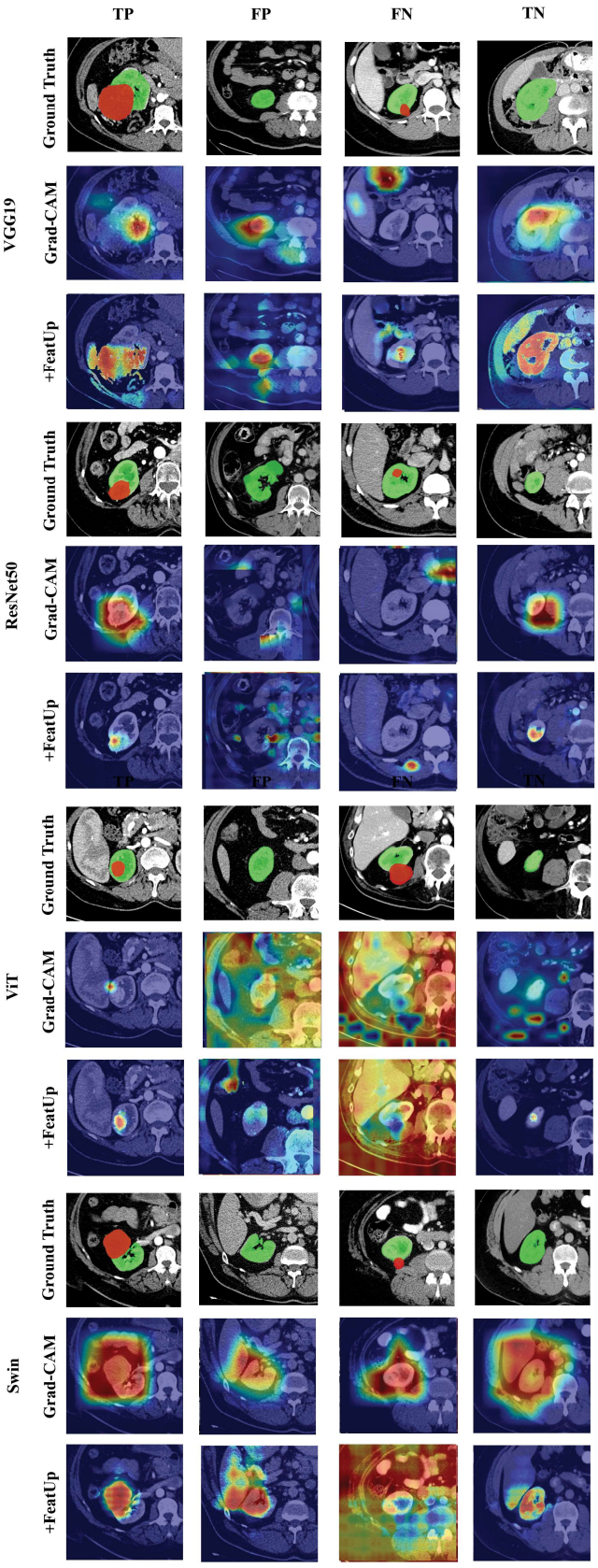

図 5　腎臓 CT におけるヒートマップ例

表 5　腎臓 CT におけるヒートマップの忠実性評価

モデル	手法	AOPC ↑	Insertion ↑
VGG19	Grad-CAM	**0.263**	0.482
	+ FeatUp	0.180	**0.541**
ResNet50	Grad-CAM	0.172	**0.829**
	+ FeatUp	**0.271**	0.790
ViT-L	Grad-CAM	**0.232**	**0.911**.
	+ FeatUp	0.123	0.883
Swin-L	Grad-CAM	**0.304**	**0.928**
	+ FeatUp	0.214	0.924

表 6　腎臓 CT におけるセグメンテーション指標評価

モデル	手法	IoU ↑	BF スコア ↑
VGG19	Grad-CAM	0.039	**0.002**
	+ FeatUp	**0.040**	0.001
ResNet50	Grad-CAM	0.191	0.072
	+ FeatUp	**0.215**	**0.122**
ViT-L	Grad-CAM	0.023	**0.016**
	+ FeatUp	**0.040**	0.013
Swin-L	Grad-CAM	0.120	0.022
	+ FeatUp	**0.134**	**0.064**

を特定するだけで確信度を急激に高めやすいドメイン特性を反映しているといえる．

6.3　セグメンテーション指標による評価

表 6 は，KiTS23 のマスク画像を正解領域としたセグメンテーション指標の評価結果である．IoU については，検証した全てのモデルにおいて提案手法がベースラインを上回った．特に ResNet50（0.191 → 0.215）および Swin-L（0.120 → 0.134）において大きな向上が確認された．また，ViT-L においても 0.023 から 0.040 へと数値が改善している．BF スコアに関しては，ResNet50 において 0.072 から 0.122 へ，Swin-L において 0.022 から 0.064 へと，精度の向上が記録された．特に Swin-L における BF スコア はベースラインの約 2.9 倍に達しており，境界の再現性が改善されたことを示している．一方で，VGG19 および ViT-L においては，BF スコアの大きな向上は見られず，低い数値に留まった．

ImageNet における物体は画像内の中央に位置し，比較的大きな面積を占める傾向があるのに対し，腎臓 CT における腫瘍は画像全体に対して微細であり，かつ周囲組織（脂肪や他臓器）と隣接している．この空間的特性の違いが定量指標の挙動の差を生んでいると考える．正解領域が小さい場合における IoU の計算感度は高く，不安定になりやすい．一方で，境界の再現性を示す BF スコアに関しては，両データセットにおいて共通して向上が確認された．この結果は，対象物の大きさやドメインに関わらず，FeatUp が物体の輪郭を捉え，視認性を向上させるという点において一貫した性能向上をもたらすことを示唆している．

7　おわりに

本研究では，深層学習モデルの判断根拠をピクセルレベルで精密に可視化することを目的に，Implicit 表現に基づく特徴量アップサンプリング手法である FeatUp と Grad-CAM を統合した高解像度可視化フレームワークを提案した．実験では，ImageNet および腎臓 CT（KiTS23）の 2 種類のデータセットを用い，CNN（VGG19, ResNet50）および Vision Transformer（ViT, Swin Transformer）の異なるアーキテクチャに対する可視化性能を多角的に評価した．評価には，定性的な視認性の比較に加え，局在化精度（IoU, BF スコア）およびモデルへの忠実性（AOPC, Insertion スコア）の計 4 種類の定量指標を用いた．

その結果，ImageNet の実験では，提案手法が全てのモデルにおいて BF スコアや IoU を一貫して向上させ，視覚的な説明性の向上において有効性を示した．また，医用画像ドメインの腎臓 CT においても，従来の Grad-CAM で見られた周囲組織へのにじみを抑制した，シャープなヒートマップ生成が可能であることを確認した．一方で，忠実性評価では従来手法に比べて性能が低下することも確認された．

今後の課題として，Grad-CAM と自己注意機構の不整合が示唆されたため，今後は Attention Rollout [1] や Transformer Explainability [5] 等の手法に FeatUp を適用し，アーキテクチャの特性を反映した高解像度化アプローチを模索する必要がある．また，Grad-CAM++ [4] や Score-CAM [20] といった Grad-CAM の派生手法への拡張も考えられる．

謝　辞

本研究は JSPS 科研費 25K15238 の助成を受けたものです．

文　献

[1] S. Abnar and W. Zuidema. Quantifying attention flow in Transformers. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197, 2020.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 5, pp. 898–916, 2011.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, Vol. 58, pp. 82–115, 2020.

[4] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pp. 839–847, 2018.

[5] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Int'l Conf. on Computer Vision (ICCV)*, pp. 103–112, 2021.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer,

G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the Int'l Conf. on Learning Representations (ICLR)*, 2021.

[7] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV)*, pp. 3429–3437, 2017.

[8] X. Fu, Z. Li, Y. Sun, J. Gu, A. L. Yuille, and X. Wang. FeatUp: A model-agnostic framework for features at any resolution. In *Proc. of the Int'l Conf. on Learning Representations (ICLR)*, 2024.

[9] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr. Large-scale unsupervised semantic segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 11036–11045, 2022.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[11] N. Heller, et al. The KiTS21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT, arXiv:2307.01984, 2023.

[12] K. Kadowaki and Y. Kameya. Visibility improvement in Grad-CAM via high-resolution feature maps with FeatUp. In *Proc. of The 2025 Principle and Practice of Data and Knowledge Acquisition Workshop (PKAW)*, pp. 201–212, 2025.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proc. of the IEEE/CVF Int'l Conf. on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pp. 405–421, 2020.

[15] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, pp. 62–66, 1979.

[16] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2018.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV)*, pp. 618–626, 2017.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Proc. of the Int'l Conf. on Learning Representations (ICLR)*, 2015.

[19] T. Tanaka, T. Fukazawa, Y. Kameya, K. Yamada, K. Hotta, T. Takahashi, N. Sassa, Y. Matsukawa, S. Iwano, and T. Yamamoto. Kidney cancer detection from ct images by transformer-based classifiers. In *Proc. of the IIAI Int'l Conf. on Advanced Applied Informatics (IIAI-AAI)*, pp. 456–461, 2023.

[20] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 24–25, 2020.

追加学習前後のモデルの予測確率の変化に基づく情報の特異性推定

疋田　知寛†　　湯本　高行†

† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8 – 2 – 1
E-mail: †ad25w059@guh.u-hyogo.ac.jp, ††yumoto@sis.u-hyogo.ac.jp

あらまし　インターネット上には膨大な情報が存在し，その中には信頼性に乏しいものも含まれる．たとえば，新型コロナウイルスのパンデミック時には「ワクチンにマイクロチップが含まれている」といった偽情報が拡散した．このような背景から，ユーザ自身による主体的な情報精査の重要性が高まっている．本研究では，情報の偏りを「特異性」として捉え，言語モデルを用いてその特異性を推定することで，ユーザによる判断を支援する枠組みの構築を目的とする．具体的には，汎用的な言語モデルと，偏った情報で追加学習したモデルの単語予測確率の差を利用し，文章全体およびキーフレーズの特異性を推定する．偏った情報と一般的な情報を用いた実験により，手法の有効性を検証する．

キーワード　有害情報・偽情報・誤情報の検出と排除，LLM

1 はじめに

現在，人々はインターネット上の多くの情報に接しているが，その中には信頼性に問題のある情報も含まれており，玉石混合の状態である．そのため，情報の精査の必要性は高まっているが，総務省が 2025 年 3 月 31 日から 4 月 2 日にかけて全国の 15 歳以上の男女 2,820 人を対象として実施した ICT リテラシー実態調査 [1] によると，偽・誤情報に接触した人のうち約 25 パーセントが何らかの手段で情報を拡散したこと，回答者の約 90 パーセントが ICT リテラシーが重要と考える一方，回答者の約 70 パーセントは ICT リテラシー向上に向けた具体的な取り組みを行っていないことが明らかになった．さらに，取り組みを行っていない理由として最も多かった回答は「取組み方が分からないから」であった．このようにユーザの多くが ICT リテラシーが重要であるという認識を持ちながら，具体的な方法が分からないために情報の精査を行うユーザが少ないという現状が考えられる．

そこで，本研究では情報の特異性という指標をユーザに提示することで，ユーザ自身による情報の精査を支援する枠組みを最終的な目的とし，その基盤となる技術として，本論文では文章中に含まれる情報の偏りを定量的に評価する手法を提案する．本研究における特異性とは，対象となる文章が，あるトピックにおける一般的な文脈からどの程度逸脱しているかを示す指標である．具体的には，事前学習済みの言語モデルと，偏った情報を用いて追加学習した言語モデルの予測確率の差に着目し，両者の差を KL ダイバージェンスとして定量化することで，文章中の各要素における特異性を推定する．このとき，文章全体に対する特異性に加えて，文章のキーフレーズに基づく特異性の評価を行う．

本研究の最終的な目的は，このようにして得られた特異性をユーザに分かりやすく可視化し，情報の精査を促すことである．本論文ではその第一段階として，文章およびキーフレーズに対する特異性推定手法の構築と評価を行う．可視化手法の設計およびユーザインタフェースへの実装及びその評価については，今後の課題とする．

本手法は，文章が持つ偏りの程度を特異性として定量的に推定することで，ユーザが主体的に情報を再考するきっかけを与える情報の定量化を目指す．

2 関連研究

現代の情報環境においては，様々な発信者による多様な情報が流通している一方で偽・誤情報や，陰謀論と呼ばれる情報など，偏りのある情報が広く拡散される問題が指摘されている．このような情報について様々な研究が行われてきた．本章では，このような偏りのある情報に関する研究を中心に紹介し，本研究の位置づけを明らかにする．

インターネット上における偏った情報や偽情報の拡散は，近年社会的に大きな問題となっており，これに伴い偏った情報の内容的特徴や拡散構造を明らかにする研究が国内外で進められてきた．

栗原の研究では，コロナ禍における日本語書籍や言説を対象に，陰謀論的主張やワクチン否定的言説の広がりを分析している [2]．この研究では，特定の出版社や著者により，反ワクチン的立場を取る書籍が集中的に刊行されていたことが示されており，情報発信者と情報の偏りの関係性が指摘されている．

鳥海らは，日本語 Twitter データを用いて反ワクチン言説の拡散構造と，その社会的・政治的背景を分析している [3]．同研究では，反ワクチン的な情報を発信・拡散するアカウントが，必ずしも一様な集団ではなく，陰謀論，スピリチュアリティなど，複数の関心領域と結びついて形成されていることが示されている．また，これらの言説はソーシャルメディア上でクラスターを形成し，特定の発信者や影響力のあるアカウントを中心として拡散する傾向があることが報告されている．さらに，反ワクチン的な言説は時間の経過とともに新たな層を取り込みながら拡大しており，特にパンデミック以降に反ワクチン的態度を形成した利用者は、健康やスピリチュアルな関心を入口として当

該言説に接触している可能性が示唆されている．このように，偏った情報はエコーチェンバー的な構造や発信者ネットワークと密接に関係しており，単なる内容分析にとどまらず，言説の特徴や文脈を考慮した分析が重要であると考えられる．

Lai らの研究では，Twitter 上での偽・誤情報の拡散状況を，6 つの言語（英語、日本語、スペイン語、フランス語、ドイツ語、韓国語）で分析している [4]．その結果，Twitter 上で最大のユーザー数を持つ英語と比較しても、特定のトピックにおいては、日本語でのリツイート数が英語を上回るなど、日本が誤情報の普及率が低い国であるという従来の定説に疑問を投げかける結果となった．

本研究では，個々の文章がどの程度一般的な文脈から逸脱しているかを定量的に評価する手法を提案する．文章単位の入力を対象とし，汎用言語モデルと偏った情報で追加学習した言語モデルの予測差に基づいて，情報の「特異性」を推定する点に特徴がある．これは，情報の真偽を判定するファクトチェックとは異なり，文章が持つ偏りや異質性の度合いを可視化することで，ユーザ自身による情報精査行動を支援する枠組みを目指すものである．

3 提 案 手 法

本研究では，情報の偏りを評価する指標として，特異性という独自の指標を用いる．偏った情報とは，あるトピックにおいて特定の主張や立場に基づいて発信されている情報のことを言い，そのような情報には偽・誤情報などが含まれる場合が考えられる．このような情報は，一般的な情報と比較して特異な語や文脈を含む場合がある．このような，ある文章や語句が，特定のトピックにおける一般的な文脈からどの程度異質であるかを示す指標として特異性を用いる．つまり，あるトピックにおいて多数の主流な情報に基づいて形成される語彙分布や文脈構造を一般的な情報としたとき，これに対して偏った情報では，一般的な情報には見られない語句や表現が見られる場合が考えられる．このような異質さの度合いを示す指標を特異性とする．本研究では，特異性を言語モデルの予測確率分布の差として捉える．同一の文章に対して，学習前の汎用モデルと，偏った情報を用いて追加学習したモデルとでマスク予測を行い，その予測確率の分布の差を比較することで，特異性の定量化を行う．

3.1 提案手法の概要

まず，入力文章よりキーフレーズを抽出する．キーフレーズの抽出には既存の大規模言語モデル（LLM）を用い，文章中に実際に出現する語句の中から，重要な語を抽出する．

次に，特異性推定のために，汎用的な事前学習済み BERT モデルと，偏った情報を含むコーパスで追加学習した biased BERT の 2 つのモデルを用いる．文章全体の特異性は，各トークンを順次マスクした際の両モデルの予測確率分布の差を KL ダイバージェンスとして算出し，その平均値により評価する．

さらに，キーフレーズに対応する箇所についても同様にマスク予測を行い，当該箇所における KL ダイバージェンスを算出することで，重要語単位の特異性を推定する．これにより，文章全体の特異性と，その根拠となるキーフレーズの特異性を推定する．

3.2 BERT による Masked Language Modeling

本研究では，BERT（Bidirectional Encoder Representations from Transformers）[5] の事前学習済みモデルを基盤として使用する．BERT は，Transformer のエンコーダ部分を多層に積み重ねた構造を持ち，大規模なテキストコーパスを用いた事前学習により，文脈を考慮した単語の分散表現を獲得している．BERT の事前学習手法の一つである Masked Language Modeling（MLM）では，入力文中の一部のトークンをマスクトークン [MASK] に置き換え，その位置の元の単語を予測するタスクを通じて学習が行われる．具体的には，入力文 $x = (x_1, x_2, \ldots, x_n)$, に対し，ある位置 i のトークン x_i を [MASK] に置換した入力 $\tilde{x}^{(i)}$ をモデルに与え、x_i の予測確率分布 $P(x_i|\tilde{x}^{(i)})$ を出力する．

本研究では，この MLM の仕組みを利用し，同一の入力文に対して汎用的な事前学習済みモデルと，3.3 節で後述の，偏った情報で追加学習したモデル（biased BERT）の予測確率分布を比較する．これにより，追加学習によってモデルの予測がどのように変化したかを定量的に捉える．

3.3 biased BERT

本研究では，汎用的な日本語 BERT モデルを基盤とし，偏った情報を含むテキストデータを用いて MLM による追加学習を行ったモデルを biased BERT と定義する．追加学習に用いるデータは，特定の主張や立場に偏った内容を含む文章群であり，これによりモデルは，そのような文脈に特徴的な語彙や表現に対して高い確率を割り当てるようになると考えられる．

biased BERT は，偏った情報の分布を内部表現として獲得していると考えられる．本研究では，この biased BERT と学習前のモデルとの予測確率の差を利用することで，特異性の推定を行う．

3.4 キーフレーズ抽出

本研究では，キーフレーズを，文章の主要な内容を構成する語または語句と定義する．具体的には，文章中に実際に出現する名詞句・動詞句等のうち，その文章の意味を理解する上で重要な役割を果たすものを指す．キーフレーズの抽出には，LLM を用い，定義したキーフレーズを抽出するようにプロンプトで指示を与え，文章とともに入力し，出力で得た語句をキーフレーズとする．想定されるキーフレーズ抽出結果の具体例を表 1 に示す．

3.5 KL ダイバージェンスに基づく特異性推定

特異性の推定は，文章全体および抽出されたキーフレーズに対して行う．文章全体の特異度については，各トークンを順次マスクし，biased BERT と，その学習前の基のモデルそれぞれのマスク予測確率分布を算出する．その分布間の差を KL ダイバージェンスとして求め，全トークンにおける平均値を文章

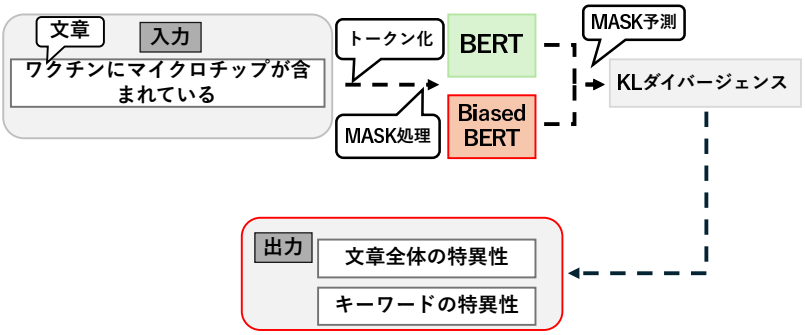

図 1　特異性推定のイメージ

表 1　キーフレーズ抽出の具体例

例	入力文章	抽出されたキーフレーズ
1	私たちが目撃しているのは単なる医療過誤ではなく，mRNA ワクチンを通じた人類への組織的な攻撃なのだ.	mRNA ワクチン，組織的な攻撃
2	厚生労働省は感染拡大を防ぐため，ワクチン接種の有効性を示すデータを公表した.	防ぐ，ワクチン接種，有効性，公表
3	この博物館は 19 世紀に建設され，現在も当時の建築様式を残している.	博物館，当時，残し

全体の特異度と定義する．一方，キーフレーズに対する特異度については，抽出されたキーフレーズに対応する位置をマスクし，同様に KL ダイバージェンスを算出する．このように，文章全体の特異性と，その根拠となる語句の特異性を推定する．この手法のイメージを図 1 に示す．

本手法では，2 つのモデルの予測確率の分布の差異を測る指標として，KL ダイバージェンスを用いる．確率分布 P と Q に対する KL ダイバージェンスは，以下の式で定義される．

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{1}$$

KL ダイバージェンスは，確率分布 P から見て Q がどの程度異なるかを表す非対称な指標であり，P と Q が一致する場合に 0 となり，分布の差異が大きいほど値が大きくなる．本研究では，ある位置 i をマスクした際の，基のモデルの予測確率分布 $P_{base}(x_i|\tilde{x}^{(i)})$ と biased BERT の予測確率分布 $P_{biased}(x_i|\tilde{x}^{(i)})$ との KL ダイバージェンスを算出する．また，KL ダイバージェンスの計算において，BERT のトークナイザーの全語彙（約 32,000 語）に対してではなく，2 つのモデルの予測確率が高い上位 5 語彙に限定して算出を行う．具体的には，あるトークン位置 i をマスクした際に，biased BERT の予測分布から確率上位 5 トークン，基のモデルの予測分布から確率上位 5 トークンをそれぞれ抽出し，両者を合わせたトークン集合 $V_{top}^{(i)}$ を構成する．次に，$V_{top}^{(i)}$ に含まれる各トークンの確率値について，その合計が 1 になるように正規化を行う．このとき，2 つのモデルの予測確率分布における KL ダイバージェンス $D_{KL}^{(i)}$ は次の式で表すことができる．

$$D_{KL}^{(i)} = \sum_{j \in V_{top}^{(i)}} \tilde{P}_{biased}^{(i)}(j) \log \frac{\tilde{P}_{biased}^{(i)}(j)}{\tilde{P}_{base}^{(i)}(j)} \tag{2}$$

j は $V_{top}^{(i)}$ に含まれる各トークン，$\tilde{P}_{biased}^{(i)}(j)$ は biased BERT がトークン j を予測する正規化後の確率を表し，$\tilde{P}_{base}^{(i)}(j)$ は基のモデルがトークン j を予測する正規化後の確率を表す．また，$D_{KL}^{(i)}$ の値が大きいほど，偏った情報による追加学習の影響が強く表れていることを意味し，当該トークンまたはその文脈が偏った情報に特徴的である，すなわち特異度が高いと解釈する．文章全体の特異度は，全トークン位置における KL ダイバージェンスの平均値として算出する．

$$S_{text}(x) = \frac{1}{n} \sum_{i=1}^{n} D_{KL}^{(i)} \tag{3}$$

ここで，$S_{text}(x)$ は文章 x 全体の特異度を表す．また，キーフレーズに基づく特異度は，抽出されたキーフレーズに対応する位置の KL ダイバージェンスとして評価する．キーフレーズ k が位置 j から $j+m-1$ までの m 個のトークンから構成される場合，そのキーフレーズの特異度 $S_{keyphrase}$ は以下のように算出する．

$$S_{keyphrase}(k) = \frac{1}{m} \sum_{i=j}^{j+m-1} D_{KL}^{(i)} \tag{4}$$

また，文章から複数のキーフレーズが抽出される場合は，それらの平均値（mean 集約）および最大値（max 集約）を用いて評価を行う．

4　実　　験

本章では，提案手法に基づくキーフレーズ抽出および特異性推定の実験結果について述べる．本研究では，提案手法のキーフレーズ抽出部分および特異性推定部分の評価実験を行った．

4.1　キーフレーズ抽出実験

本実験では，LLM を用いたキーフレーズ抽出手法の有効性を検証する．具体的には，文章からキーフレーズを適切に抽出できるか，また抽出されたキーフレーズがカテゴリによってどのような傾向を示すかを分析する．

4.1.1　実 験 手 法

実験には，コロナウイルスに関する 3 カテゴリのテキスト

データを用いた．各カテゴリ 30 文，計 90 文を用意した．

- 偽情報：コロナワクチンの危険性を誇張する主張など，偏った情報を発信しているウェブサイトから収集したコロナワクチンに関する文章
- 正情報：医療機関等の公式サイトから収集したコロナワクチンに関する一般的な情報
- 無関係：コロナウイルスとは無関係な分野（歴史，科学，文化など）の文章

各文章は句点で区切られた 1 文を基本単位とし，実験データとして使用した．

キーフレーズ抽出には，ELYZA 社が公開している日本語 LLM の elyza/Llama-3-ELYZA-JP-8B[1]を使用した．入力として 1 文を与え，事前に設計したプロンプトに従って語句の抽出を行う．プロンプトは，具体的なタスク内容と制約条件を明示した．また，few-shot プロンプティング手法を用いて，3 つの具体例を提示することで，期待する出力形式を与えた．実際のプロンプトを以下に示す．

プロンプト：

次の文章からキーフレーズとなる必要十分の語句を抜き出し，それのみをカンマ区切りで出力せよ．説明や前置きはしないこと．

条件#
- キーワードは必ず本文中にそのまま出現する語句のみ（言い換え・創作・一般化は禁止）
- 助詞・助動詞などの機能語だけは不可（例：「は」「が」「する」だけ等）
- 出力は行のみ，半角カンマ 1(,)区切り
- 説明文・前置き・箇条書き・引用符は禁止
- 指定条件以外の出力は禁止

例 1文章：私たちが目撃しているのは単なる医療過誤ではなく、ワクチンを通じた人類への組織的な攻撃なのだ．
「ワクチン，組織的な攻撃」

例 2文章：厚生労働省は感染拡大を防ぐため、ワクチン接種の有効性を示すデータを公表した．「防ぐ，ワクチン接種，有効性，公表」

例 3文章：この博物館は１９世紀に建設され、現在も当時の建築様式を残している．「博物館，当時，残し」

プロンプトでは，以下の制約を与えた．

- 本文中の語句のみを対象とすること
- 機能語のみからなる語句を除外すること
- 説明文や前置きを出力しないこと

1：https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B

これにより，モデルによる言い換えを抑制し，文章中の語句を抽出する設定とした．

4.1.2 実験結果

抽出されたキーフレーズの数は文ごとに異なり，1 文あたり 1〜4 語の範囲に分布した．表 2 に，各カテゴリから抽出されたキーフレーズの例を示す．

表 2 各カテゴリにおけるキーフレーズ抽出結果の例

カテゴリ	文章例	抽出されたキーフレーズ
無関係	高市総理は 14 日，参議院・予算委員会に出席し…	参議院，高市総理，ガス料金，予算委員会
正情報	人の体には，もともと病原体に対する免疫力が備わっています	人，体，免疫力
偽情報	ワクチンの害がシェディングという形でどんどんと…	ワクチン，害，シェディング

無関係文においては，名詞や一般的な説明語がキーフレーズとして抽出される傾向が見られた．例えば，「参議院」「予算委員会」「ガス料金」などの名詞が抽出された．正情報文では，「病原体」「免疫力」「ワクチン」など，医学的な用語が抽出された．偽情報文では，特定の主張や評価を強く表す語句がキーフレーズとして抽出される場合が多く，語彙の選択に一定の偏りが生じやすい傾向が見られた．「mRNA ワクチン」「シェディング」など，ワクチンに関する特定の主張を含む語句が多く抽出された．

4.1.3 考察

キーフレーズ抽出実験の結果から，多くの場合に名詞が抽出されていることが確認された．特に，偽情報においては特定の名詞が明確に抽出される傾向があった．

一方で，いくつかの課題も明らかになった．まず，抽出されるキーフレーズの多くが名詞であり，動詞や形容詞などの述語的要素があまり抽出されない傾向が見られた．例えば，表 2 に示すように，「参議院」「高市総理」「ガス料金」といった名詞的要素が中心となり，文章の動作や状態を表す語句の抽出が限定的であった．これは，LLM が固有名詞や専門用語をキーフレーズとして認識しやすい可能性が考えられる．今後，動詞や形容詞を含む多様な品詞のキーフレーズを抽出するためには，プロンプトの改善や後処理における品詞バランスの調整が必要である．

また，抽出されるキーフレーズ数にばらつきが存在した．これには，文章の長さや内容の複雑さに依存する傾向があると考えられるが，文章の長さにかかわらず適切なキーフレーズを抽出するようにプロンプトを改善する必要がある．

以上の課題を踏まえると，プロンプトの改善により，より適切なキーフレーズ抽出が可能にすることで，結果として特異性推定の精度向上につながることが期待される．

4.2 特異性推定の実験

4.2.1 実験手法

本実験では，4.1 節で抽出したキーフレーズおよび文章全体

に対して特異性推定を行った．偏った情報で追加学習するためのデータとして，コロナウイルスに関する偽情報 100 件を用意した．これらは医療関係の偽・誤情報を発信しているサイトから収集した，コロナウイルスやワクチンについての内容の文章である．句点で区切られた 1 文を 1 データとする．biased BERT の学習には，東北大学が公開している日本語 BERT モデル tohoku-nlp/bert-base-japanese-v3[2]を基盤のモデルとして使用した．学習手法として MLM タスクを用い，データを 3 分割し（訓練データ 66 %，検証データ 33 %），それぞれ 30 エポックずつ 3-fold の交差検証を行った．3 回の学習の中から，検証損失が最も低いエポックのモデルを本実験で用いる biased BERT として採用した．そして，基の BERT モデルと biased BERT の予測確率分布の差を KL ダイバージェンスとして算出した．

文章全体の特異度は，文中の全トークンを順次マスクした際に得られる KL ダイバージェンスの平均値として算出した．キーフレーズに基づく特異度については，4.1 節で抽出された各キーフレーズをマスクした際の KL ダイバージェンスを算出した．全ての場合において、偽情報の特異度が正情報及び無関係と比較して相対的に高くなっている状態を理想とする．

4. 2. 2　実 験 結 果

図 2 に，各カテゴリにおける文章全体の特異度の分布を示す．箱ひげ図により，中央値，四分位範囲，および外れ値を可視化した．

偽情報の特異度の中央値は約 0.61 であり，正情報（中央値約 0.30）および無関係（中央値約 0.14）と比較して明確に高い値を示した．この結果は，偽情報が一般的な文脈から逸脱している度合いが大きいことを示唆している．一方で，偽情報の中には特異性が低く評価されたものも存在した．例えば，「mRNA ワクチンは有効率が高く，重篤な有害事象が必ずしも多いわけではない」という文章は，正情報であるにもかかわらず，偽情報を否定するような内容であったため特異度が高く評価された（約 1.67）．このように，文章の内容によって特異度の評価にばらつきが見られた．また，正情報の特異性が無関係よりもやや高い傾向が見られた．これは，正情報がコロナウイルスやワクチンに関する専門用語を多く含むため，biased BERT の学習データと語彙的な重複があることが影響していると考えられる．

図 3 に，mean 集約によるキーフレーズに基づく特異度分布を示す．偽情報の特異度中央値は約 1.65 であり，正情報（中央値約 0.50）および無関係（中央値約 0.26）と比較して顕著に高い値を示した．この結果は，偽情報に含まれるキーフレーズが，一般的な文脈において特徴的であることを示している．偽情報のキーフレーズとしては，「弱毒化」「無毒化」「ウイルス」「生ワクチン」などが抽出され，これらは高い特異度を示した．一方，無関係のカテゴリでは，「占める」「7 番目」「含まれる」「カーナライト」といったキーフレーズが比較的高い特異度（約 0.83）を示す例外も見られた． 図 4 に，max 集約によるキーフレーズに基づく特異度の分布を示す．偽情報の特異度の中央値は約 2.86 であり，正情報（中央値約 1.03）および無関係（中央値約 0.56）と比較して非常に高い値を示した．mean 集約の場合と比較して，max 集約の場合では偽情報と他のカテゴリとの差がより明確になった．これは，偽情報に含まれるキーフレーズの中に，特に特異度の高いものが含まれる傾向があることを示していると考えられる．max 集約を用いることで，文章中の最も特徴的な語句に基づいて特異性を評価できる利点があると考える．

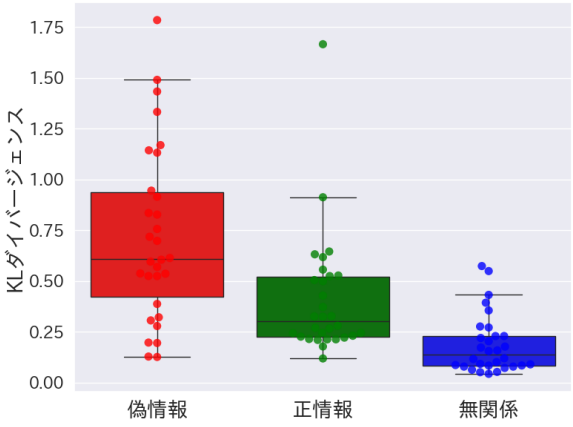

図 2　文章全体の特異度の分布

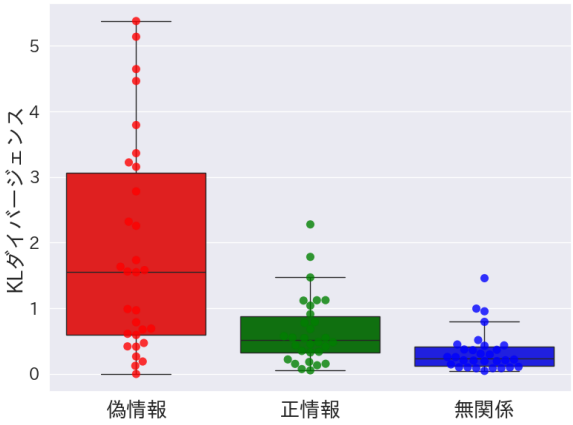

図 3　キーフレーズに基づく特異度の分布（mean 集約）

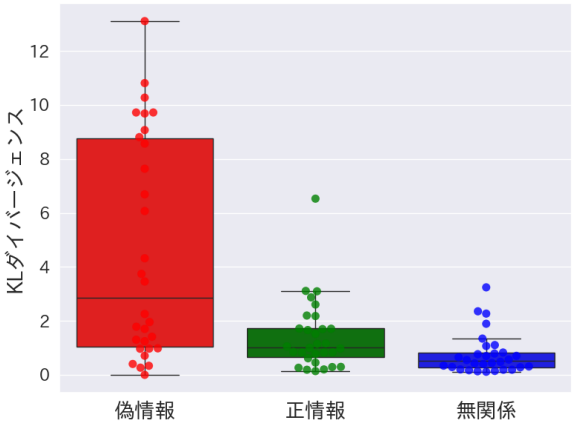

図 4　キーフレーズに基づく特異度の分布（max 集約）

2：https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

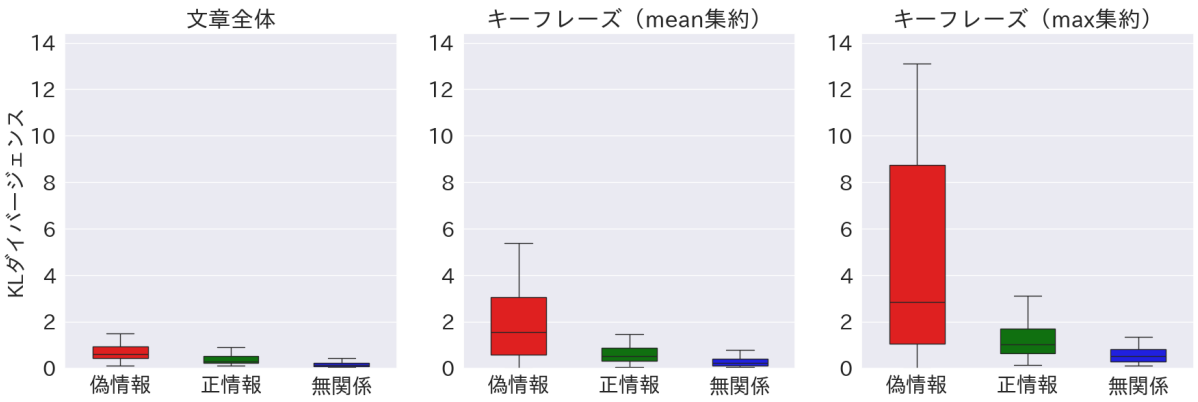

図 5　同一の y 軸のスケールによる 3 つの指標の特異度の分布

図 2〜図 4 では，3 つの指標（文章全体，キーフレーズ mean 集約，キーフレーズ max 集約）の分布を詳細に観察するため，y 軸のスケールを個別に設定している．3 つの指標の分布について，y 軸のスケールを統一したを図 5 に示す．

図 5 より，y 軸のスケールを統一して比較すると，キーフレーズに基づく特異度は，文章全体と比較して偽情報の特異度が高い値であることが確認できる．これは，キーフレーズに焦点を当てることで特異性の差がより顕著になることを示唆していると考える．文章全体の特異度はキーフレーズによる指標と比較して低くなっている．これは，文章全体の特異度を全トークンの平均で求めるため，一般的な語句も考慮されることから特異度が希釈されているためと考えられる．

4.3　考　　察

実験結果から，提案手法は偽情報と正情報・無関係を区別する上で一定の有効性を示すことが確認された．特に，キーフレーズに基づく特異度（mean 集約および max 集約）は，文章全体の特異度と比較して，偽情報を区別できる傾向が見られた．文章全体の特異度では，偽情報の中央値が正情報や無関係よりも高い傾向が見られたものの，分布に重なりがあった．これは，文章全体を平均化することで，特徴的な語句の影響が希釈されるためと考えられる．一方，キーフレーズに基づく特異性では，4.1 節で抽出した文章のキーフレーズとなる語句に焦点を当てることで，より明確な区別が可能となった．

本研究にはいくつかの課題が存在する．まず，正情報の特異性が無関係よりもやや高い傾向が見られた点である．これは，正情報と biased BERT の学習データが同じトピック（コロナウイルス）を扱っているため，特に偽情報を否定するような文脈の場合において語彙的な重複が生じたことが原因と考えられる．この問題に対処するためには，間違った情報を否定するような文脈の情報には評価を逆転させるといった対策が必要であることが示唆された．また，一部の偽情報で特異性が低く評価される例が見られた．これは，一般的な語彙で記述された偽情報が含まれていたためと考えられる．このことから，今後は文脈情報をより詳細に考慮した特異性推定手法の検討を行う必要がある．

5　おわりに

本研究では，情報の偏りを特異性として定量的に評価する手法を提案した．具体的には，汎用的な事前学習済み言語モデルと，偏った情報で追加学習したモデルの予測確率の差を KL ダイバージェンスとして算出することで，文章全体およびキーフレーズに基づく特異性を推定する手法を構築した．

実験では，コロナウイルスに関する偽情報，正情報，無関係の 3 種類のテキストデータを用いて，提案手法の評価実験を行った．キーフレーズ抽出実験では，LLM に対して文章のキーフレーズを抽出するようプロンプトを与えた．その結果，偽情報においては，特定の語彙が明確に抽出される傾向が見られた．特異性推定実験では，偽情報は正情報や無関係と比較して高い特異度を示すことが確認された．特に，キーフレーズに基づく特異度（mean 集約および max 集約）は，文章全体の特異度と比較して，偽情報をより明確に識別できることが示された．これは，キーフレーズの抽出によって，LLM が特徴的な語句を抽出したためと考えられる．

一方で，いくつかの課題も明らかになった．まず，抽出されるキーフレーズの多くが名詞であり，動詞や形容詞などの述語があまり抽出されない傾向が見られた．これは，プロンプト設計が主な原因と考えられる．また，それぞれのカテゴリにおいて特異性の推定が理想的な結果とならなかった例が確認された．これは，偽情報を否定する内容や，一般的な表現で記述された情報が含まれているなど文法・文脈的な要因が考えられる．

今後の課題として，以下の点が挙げられる．まず，文法的により妥当なキーフレーズ抽出を行い，特異性推定の精度を向上させる必要がある．さらに，文脈情報をより詳細に考慮した特異性推定手法の開発も重要である．実験では，LLM に対して few-shot プロンプティングを用いて文章から直接キーフレーズを抽出する手法を採用したが，抽出されるキーフレーズの多くが名詞に偏る傾向が見られた．この課題に対処するため，今後は LLM による要約の生成を経由したキーフレーズ抽出手法を検討する．実験では，偽情報を否定する内容を含む文章が高い特異性を示すという問題が確認された．これは，biased BERT が偽情報のトピックに関連する語彙全般に高い確率を割り当て

るためであると考えられる．今後は，係り受け解析を用いて否定表現を検出し，否定文脈を含む文章については特異性の評価を逆転させるなどの対策を検討する必要がある．本研究の最終的な目的は，推定された特異性をユーザに分かりやすく可視化し，情報の精査を促すことである．本論文では，特異性推定手法の構築と評価を行ったが，本手法の改善及び可視化手法の設計，ユーザインタフェースへの実装については今後の課題とする．特異性という指標をユーザに提示することで，ユーザ自身による主体的な情報精査を支援する実用的なシステムの実現を目指す．

謝　　辞

本研究は JSPS 科学研究費助成事業 24K15195 による助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] 総務省．ICT リテラシー実態調査．Technical report, 総務省, 2025. 2025 年 3 月 31 日-4 月 2 日実施.

[2] 栗原健太．コロナ禍における日本の陰謀論を問う．日本文化論年報, Vol. 26, pp. 127–179, 2023.

[3] Fujio Toriumi, Takeshi Sakaki, Tetsuro Kobayashi, and Mitsuo Yoshida. Anti-vaccine rabbit hole leads to political representation: the case of twitter in japan. *Journal of Computational Social Science*, Vol. 7, No. 1, pp. 405–423, 2024.

[4] Cameron Lai, Fujio Toriumi, and Mitsuo Yoshida. A multilingual analysis of pro russian misinformation on twitter during the russian invasion of ukraine. *Scientific Reports*, Vol. 14, No. 1, p. 10155, 2024.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

次元削減を用いたモデル非依存SHAP近似における計算効率と説明誤差のトレードオフ分析

西条　啓佑†　　宮森　　恒†

† 京都産業大学情報理工学部宮森研究室　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2253578,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では, 次元削減を用いたモデル非依存 SHAP 近似における計算効率と説明誤差の分析に取り組む. 説明可能 AI(eXplainable AI, XAI) の代表的手法である SHAP は, モデル非依存性と理論的妥当性を備える一方で, 特徴量数に対して指数的な計算量を要するため, 高次元データへの適用が困難である. 既存の高速化手法の多くは特定のモデル構造に依存しており, モデル非依存性と計算効率の両立は依然として課題である. 本稿では, 次元削減を前処理として導入した SHAP 近似手法に着目し, 計算効率と説明誤差のトレードオフを体系的に分析する. 具体的には, 線形および非線形の次元削減手法を用いて特徴空間を低次元化した後,Kernel SHAP を適用し, 元特徴空間における Shapley 値近似を行う. この際, 次元削減により生じる説明誤差を, 値誤差および特徴重要度ランキングの観点から定量的に評価する. 実験では, 複数のデータセットおよび次元削減設定に対して, 計算時間と説明誤差の関係を比較し, 次元削減を用いたモデル非依存 SHAP 近似の有効性と限界について考察する.

キーワード　XAI, SHAP, 次元削減, モデル非依存, 計算効率, 誤差分析

1　はじめに

近年,AI 技術の急速な発展により, 深層学習をはじめとする高性能な機械学習モデルが広く利用されている一方で, これらのモデルは高い非線形性や多層構造を持つため, モデル内部の予測過程が人間にとって理解困難となるブラックボックス化の問題が指摘されている [1], [2].AI 技術の社会実装において, 単に高い予測品質を達成するだけでなく, モデルがどのような根拠に基づいて予測を行なっているのかを人間が理解可能な形で説明できることが重要視されており,AI の振る舞いや判断の根拠を可視化することで, モデルの信頼性や妥当性を裏付ける技術を説明可能 AI(eXplainable AI, XAI) と呼ぶ [2], [3].

XAI には様々な手法が提案されているが, その中でも SHAP(SHapley Additive exPlanations) [6] は代表的な手法の一つである.SHAP は, ゲーム理論における Shapley 値の概念を応用することで, 各入力特徴量が予測に与える寄与度を定量的に評価する手法であり, 一貫性や公平性が理論的に保証されている. また, モデル非依存性を持つため, 様々な機械学習モデルに適用可能であり,XAI のデファクトスタンダードとして広く認知されている. 一方で,SHAP には計算コストが非常に高いという課題が存在する. これは,SHAP が各特徴量の寄与度を算出する際に, 全ての特徴量の部分集合を考慮してモデルを評価する必要があり, 特徴量数の増加に伴って計算量が指数的に増加するためである. そのため, 高次元データや大規模モデルに対しては SHAP をそのまま適用することが困難であり, 計算効率の改善が重要な課題となっている.

この課題を解決するため, これまでに様々な高速化手法が提案されている. これらの手法は, モデルの構造を利用するモデル依存型手法と, モデルの種類に依存しないモデル非依存型手法に分類される. モデル依存型手法は特定のモデル構造を活用することで非常に高い計算効率を実現できるが, 他のモデルには直接適用できないという制約を持つ. 一方, モデル非依存型手法はモデルの内部構造に依存せず様々なモデルに適用可能であるが, 構造を利用した効率化ができないため, 多数のサンプルに対してモデルを評価する必要があり, 計算コストは依然として高い. このように, 既存手法はいずれも一定の有効性を持つ一方で, 計算効率の高さとモデル非依存性を同時に達成する手法は未だ確立されていない.

モデル非依存 SHAP 近似手法の計算コストの高さは, 主に入力特徴量の次元数の高さに起因する. これは, 特徴量の部分集合に対するモデル出力を用いて寄与度を推定するため, 特徴量数が増加すると, 考慮すべき部分集合の空間が指数的に拡大し, その空間を十分に近似するには多数のサンプルが必要となるためである. したがって, このような計算コストの増加に対しては, 特徴量空間そのものの次元を低減することが有効な対策となり得る.

高次元データに対する一般的なアプローチとして, 次元削減を用いることでデータの本質的な情報を保持しつつ次元を小さくし, 機械学習モデルの計算効率や性能を向上させる先行研究が複数報告されている [4], [5]. 一方で, 次元削減を SHAP の計算過程に組み込んだ場合に, 得られる Shapley 値の説明効果に与える影響については十分な検討がされていない. 特に, どの程度の次元削減が許容されるのか, また説明誤差が計算効率とどのような関係を形成するのかについては整理されていない. 本稿はこの点に着目し, 次元削減を前処理として導入した場合のモデル非依存型 SHAP 近似手法の計算効率と説明誤差の関係を整理するとともに, その特性を明らかにする.

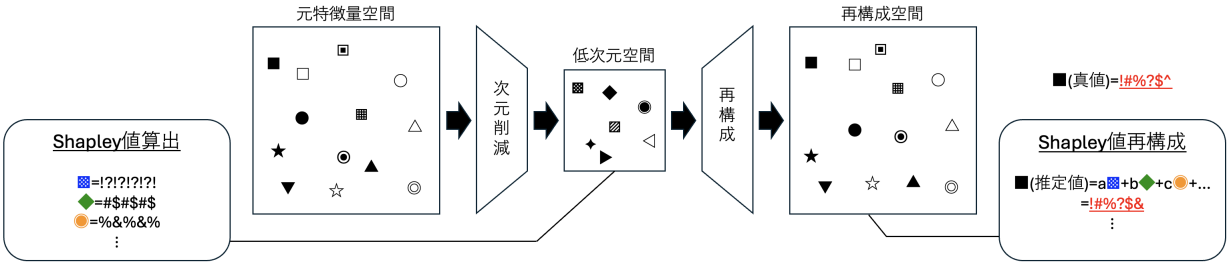

図 1　本研究の概要

本研究の目的は, 次元削減を導入したモデル非依存型 SHAP 近似手法において, 計算効率の向上と説明誤差の増加との間に生じるトレードオフを体系的に分析することである. 本研究の概要を表した図を図 1 に示す. 本研究では, 次元削減後の低次元表現に基づく Shapley 値の推定を対象とし,Shapley 値の真値と推定値を比較することで, 次元削減によって生じる説明誤差の特性や計算コストとの関係を定量的に評価する. これにより, 高次元データに対してモデル非依存性を維持したまま SHAP を適用する際の実務的な指針を与えることを目指す.

本論文の貢献は以下の通りである.

(1) 次元削減を前処理として導入したモデル非依存 SHAP 近似における計算効率と説明誤差の関係を体系的に分析した
(2) データの性質や分析の目的に応じて, 次元削減の適用の可否判断および手法選択を行うための指針を与えた

本論文の構成は以下の通りである.2 節では, 関連研究として Shapley 値算出の高速化手法を提案した研究と, 次元削減によるモデルの性能向上を示した研究と, 次元削減が SHAP の解釈性に与える影響を示した研究について紹介する.3 節では,SHAP 及び Kernel SHAP の理論的背景を整理し, 次元削減を導入した際に生じる計算効率と説明誤差のトレードオフについて述べる.4 節では,3 節で整理した理論的観点に基づいて評価実験を行い, 計算効率と説明誤差の関係を定量的に分析・考察する.5 節では, 本論文の結論と今後の課題をまとめる.

2　関 連 研 究

本節では,Shapley 値の算出や次元削減の応用に関連する研究として,2.1 節で Shapley 値算出の高速化手法を提案した研究,2.2 節で次元削減によるモデルの性能向上を示した研究を紹介した後,2.3 節で次元削減が SHAP の解釈性に与える影響を示した研究を紹介する.

2.1　Shapley 値算出の高速化手法

Shapley 値算出の高速化手法を提案した代表的な研究として,TreeSHAP を提案した研究が存在する [7].TreeSHAP は, 決定木モデルの構造を活用することで Shapley 値を厳密かつ高速に算出できるモデル依存型手法である. また, 各特徴量の寄与は木の経路で表現されるため, 特徴量間の関係が直感的に解釈しやすいという特徴がある.

また, 別の高速化手法として,Kernel SHAP も提案されている [6].Kernel SHAP は, 重み付き線形回帰を用いることで Shapley 値を近似的に推定するモデル非依存手法である. モデルの内部構造に依存しないため, 様々なモデルに適用可能であるという利点がある.

しかし, 特徴量次元そのものを削減することで, モデル非依存性を保ったまま Shapley 値算出を効率化するアプローチについては, 計算効率と説明誤差の関係が体系的に整理されていない.

2.2　次元削減によるモデルの性能向上

次元削減によるモデルの性能向上を示した研究として, 多言語 Transformer の文埋め込みに対して各種の次元削減手法を適用し, 性能を維持できることを示した研究が存在する [8]. この研究では, いくつかの手法において埋め込み次元を大幅に削減しながらも意味類似度の正確性をほぼ維持でき, 特定の条件下では元モデルを上回る結果が得られている.

また,Transformer の Adapter を低ランク表現し, ファインチューニングを効率化した研究も存在する [9]. この研究は, 通常の Adapter よりも大幅にパラメータ数を削減しながら, 元モデルに近い, あるいはそれ以上の性能を維持できることを示した.

いずれも, 次元削減による予測性能や学習効率, パラメータ効率の維持・向上に焦点を当てており, 次元削減後の特徴表現が説明手法にどのような影響を及ぼすかについては明示的に扱われていない.

2.3　次元削減が SHAP の解釈性に与える影響

次元削減が SHAP の解釈性に与える影響を示した研究として, データの潜在構造を捉えた低次元表現上で Shapley 値を算出し, それを元の特徴空間へ写像することで, 相関を含む高次元データに対する説明を可能とする手法を提案した研究が存在する [10]. この研究では, 潜在空間を介することで, 特徴量間の相関を考慮した Shapley 値の安定した推定と, 説明結果の解釈性向上が可能であることを示した.

また, 特徴量の相関を低減する次元操作を行った場合に,SHAP を含む説明手法の忠実性や安定性がどのように変化するかを定量的に評価した研究も存在する [11]. この研究では, 相関特徴量を削減することで, 特徴重要度ランキングの一貫性や説明の信頼性が向上する場合があることを示した.

これらの研究は, 低次元表現を用いた場合の SHAP を含む説明手法の解釈性や忠実性に重点を置いており,SHAP の計算効率や計算時間と説明誤差のトレードオフについては十分に議論されていない.

3 理論整理

本節では, 本論文で用いる理論的枠組みを整理する. まず 3.1 節で SHAP 及び Shapley 値の基本的な考え方を紹介し,3.2 節でモデル非依存型 SHAP 近似手法である Kernel SHAP の原理を説明する. その後,3.3 節で次元削減を導入した場合に生じる計算効率と説明誤差のトレードオフについて述べる.

3.1 SHAP

SHAP(SHapley Additive exPlanations) は, ゲーム理論における Shapley 値の概念を機械学習モデルの予測説明に応用した手法である [6]. なお,Shapley 値は, 多人数協力ゲームにおけるゲームの利益に対するプレイヤーの価値の順序を評価する指標であり [12],SHAP では, 各特徴量をプレイヤー, ゲームの利益をモデルの出力とみなすことで, 各特徴量の寄与度を定量的に評価できる.

本手法における, 各特徴量 i の Shapley 値を次の式 1 に示す. ここで,F は特徴量集合,S は特徴量集合 F の部分集合,x_S は集合 S における入力特徴の値を表し,$f_S(x_S)$ は特徴量集合 S が与えられたときのモデル出力を意味する.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{1}$$

式 1 を見ると,1 つの特徴量の Shapley 値の算出には, 特徴量 i を除く全ての部分集合 $S \cup \{i\}$ を考慮する必要があることがわかる. これにより, 特徴量数を $d = |F|$ とすると, 各特徴量に対して 2^{d-1} 個の部分集合が存在し, 全特徴量の Shapley 値を求める計算量は $O(d \cdot 2^{d-1})$ となるため, 特徴量数が増加すると計算量は指数的に増加する. そのため, 高次元データや大規模モデルに対して SHAP を直接適用することは困難である.

3.2 Kernel SHAP

Kernel SHAP は, 重み付き線形回帰によって Shapley 値を近似する手法である [6].

本手法では, 特徴量の有無だけで表すことで特徴空間を簡略化した二値ベクトル $z' \in \{0,1\}^M$ を用いて部分集合を表現する. そして, 部分集合 z' に対応する元特徴空間での入力を $h_x^{-1}(z')$ として復元し, そのときのモデル出力 $f(h_x^{-1}(z'))$ を用いて, 次の式 2 で表される重み付き最小二乗問題を最小化する説明モデル g を学習する. ここで,M は特徴量数を表す.

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} [f(h_x^{-1}(z')) - g(z')]^2 \pi_{x'}(z') \tag{2}$$

このとき, 説明モデル g が z' に対する線形モデルとなっており, その回帰係数が各特徴量の Shapley 値に対応する. なお, 各部分集合 z' に対する重みは, 次の式 3 で定義される.

$$\pi_{x'}(z') = \frac{(M-1)}{({}_M C_{|z'|})|z'|(M-|z'|)} \tag{3}$$

この重みは, 式 1 の係数を再現するように設計されており, 部分集合の大きさに応じて重要度を調整する役割を持つ.

このように,Kernel SHAP では,Shapley 値を回帰問題の解として表現できるため,SHAP のように全ての部分集合を考慮する必要がなく, 計算コストの削減が期待される. しかし, 部分集合の数は 2^M であり, 推定品質を高めるためには多くの部分集合に対するモデル出力の評価が必要であるため, 特に高次元データや大規模モデルでは依然として計算コストが大きい.

3.3 次元削減の導入

前節の通り,Kernel SHAP における計算コストの高さの原因は, 考慮すべき部分集合の数が特徴量数に対して指数的に増加することであった. そこで, 元の特徴量空間 $\mathbb{R}^M$ に対して次元削減を適用し, 低次元空間 $\mathbb{R}^{M'}, (M' < M)$ を得た上で, この低次元表現を説明対象として Kernel SHAP を適用することを考える.

このとき,Kernel SHAP で扱われる二値ベクトル $z' \in \{0,1\}^{M'}$ の次元数も M' に対応するため, 部分集合の数は $2^{M'}$ となる. したがって,$M' < M$ が成立する場合, 部分集合のサンプリング数及びモデル出力の評価回数が大幅に減少し, 計算コストは指数的に減少する.

この操作は Shapley 値の定義式そのものを変更するものではなく,Shapley 値を計算する対象である特徴量集合を低次元空間へ置き換える操作であるため,Shapley 値の定義や理論的性質を保持することができる. 一方で, 低次元表現は元の特徴量を圧縮した表現であるため, 低次元空間上で算出された Shapley 値は元の特徴量空間における Shapley 値と比較して情報損失が生じる. すなわち, 次元削減の導入によって計算効率は向上するが, 説明誤差が増加する可能性があるというトレードオフが存在する.

4 評価実験

本節では, 前節で説明した次元削減の導入の影響を検証するために行なった評価実験について述べる.4.1 節で本実験の目的を挙げ,4.2 節で本実験の方法を説明する. そして,4.3 節で実験結果についてまとめ,4.4 節及び 4.5 節で実験結果を分析する. その後,4.6 節で次元削減を適用する際の実用的な指針を示す.

4.1 実験目的

本実験の目的は, 次元削減を前処理として導入した場合におけるモデル非依存型 SHAP 近似手法の計算効率と説明誤差のトレードオフを体系的に評価することである. 本実験では, 次元削減によって得られた低次元表現を用いて Shapley 値を算出し, 計算時間がどの程度短縮されるかを測定するとともに, 得られた Shapley 値の説明誤差を値誤差及び特徴重要度ランキングによって定量的に比較することで, 次元削減が与える影響を明らかにする.

4.2 実験方法

本実験では, 以下に示す評価指標と実験条件を用いて, ベース

ライン手法と次元削減を導入した手法を比較評価する.

4.2.1 評価指標

i) 計算効率

計算効率の評価指標として,Shapley 値の算出に要した計算時間を測定する. 次元削減を導入する主目的が計算コストの削減であることから, 本指標は実験全体を通じた重要な評価軸である.

なお, 後述する学習を伴う次元削減手法については, 次元削減モデルの学習時間と Shapley 値の算出時間の両方を含めた総計算時間を評価対象とする. これにより, 前処理を含めた実運用上の計算効率を公平に比較できる.

ii) 説明誤差

ii-a) 値誤差

算出された Shapley 値の定量的な誤差を評価するため, 準真値と推定値との差に基づく値誤差を用いる. 本実験では, 主指標として正規化 L1 誤差を, 補助指標として RMAPE(Range-scaled Mean Absolute Percentage Error) を採用する.

正規化 L1 誤差は, 各特徴量における Shapley 値の絶対誤差の総和を, 準真値における Shapley 値の絶対値の総和で正規化した指標であり, 次の式 4 で定義される. なお,ϕ_i は特徴量 i の準真値,$\hat{\phi}_i$ は特徴量 i の推定値,d は特徴量数を表す.

$$NormL1 = \frac{\sum_{i=1}^{d} |\hat{\phi}_i - \phi_i|}{\sum_{i=1}^{d} |\phi_i|} \tag{4}$$

本指標は寄与の大きい特徴量の誤差を重視する性質を持つ. 実際の SHAP では, 少数の特徴量が大きな寄与を持つスパースな分布となる場合が多いため, 実用的な説明品質の評価に適している.

一方,RMAPE は各特徴量ごとの相対誤差を平均した指標であり, 次の式 5 で定義される.

$$RMAPE = \frac{1}{d} \sum_{i=1}^{d} \frac{|\hat{\phi}_i - \phi_i|}{\max(\phi) - \min(\phi)} \tag{5}$$

RMAPE は誤差をデータセットのレンジで正規化することで, データセットのスケールに依存せず比較可能であるという利点を持つ. ただし, 寄与の小さい特徴量に対しても同等の重みで誤差を評価するため, 微小な Shapley 値に乗るノイズの影響を受けやすい. このため, 本研究では補助的な評価指標として位置づける.

ii-b) 特徴重要度ランキング

Shapley 値の相対的な重要度の再現性を評価するため, 特徴重要度ランキングに基づく指標を用いる. 本実験では, 主指標として Overlap@k を, 補助指標として Spearman 順位相関係数を採用し, 準真値と推定値に基づくランキング間の一致度を評価する.

Overlap@k は, 上位 k 個の重要特徴に着目し, 準真値と推定値のランキングにおける共通要素の割合を測定する指標であり, 次の式 6 で定義される [15]. なお,$top_k(\pi_1)$ は準真値 π_1 における上位 k 特徴量集合,$top_k(\pi_2)$ は推定値 π_2 における上位 k 特徴量集合を表す.

$$Overlap@k(\pi_1, \pi_2) = \frac{|top_k(\pi_1) \cap top_k(\pi_2)|}{k} \tag{6}$$

Overlap@k は, 実務上重要となる上位特徴の再現性を直接評価できるという特徴を持つ.

また,Spearman 順位相関係数は,2 つのランキング間の単調関係を評価する指標であり, 次の式で定義される [16]. なお,r_i は準真値における特徴量 i の順位,$\hat{r}_i$ は推定値における特徴量 i の順位を表す.

$$\rho = 1 - \frac{6 \sum_{i=1}^{d} (r_i - \hat{r}_i)^2}{d(d^2 - 1)} \tag{7}$$

本指標は Shapley 値のスケールに依存せず, ランキング全体の整合性を評価できるという特徴を持つ. ただし, 数値が微小な下位特徴量の順位変動に影響を受けやすく, 本質的でない順位の入れ替わりがスコアを支配する可能性がある. したがって, 本研究では補助指標として併用する.

4.2.2 実験条件

i) 固定条件

次元削減の導入が SHAP 近似に与える影響を明確にするため, 予測モデル及び説明手法を固定する. 予測モデルには XGBoost を採用し, 全ての実験において同一の学習済みモデルに対して説明を行う. また, 説明手法にはモデル非依存型 SHAP 近似手法の代表的手法である Kernel SHAP を採用し, 次元削減を導入しない場合と導入した場合の差異を比較する. これにより, 次元削減の導入の有無及び次元削減手法の違いのみが結果に反映される.

さらに, 説明誤差の評価における基準として, 高サンプリング設定 (サンプリング数 50000) により算出した Kernel SHAP の結果を準真値として用いる.Kernel SHAP は理論的に Shapley 値を近似する手法であるが, 十分なサンプリング数を用いることで推定誤差を小さくできるため, 説明誤差を評価する際の準真値として適切であると考えられる. このとき, 準真値の生成とベースライン手法の両方に同一のアルゴリズムを用いることになるが, 前者は十分に大きなサンプリング数を用いて Shapley 値の収束解を近似することを目的としているのに対し, 後者は計算リソースを制限した現実的な設定下で Shapley 値を近似することを目的としている点に違いがある.

なお, 実験は単一の計算機環境上で実施した. 全手法は同一のハードウェア及びソフトウェア環境において実行されており, 計算時間に関する比較が公平に行われるよう配慮している. 実装の詳細については付録 A に示す.

ii) 変更条件

ii-a) 次元削減手法

次元削減の導入による影響を, その設計思想の観点からより多角的に検証するため, 次元削減手法および各手法のパラメータを変更する. 本研究では, 標準的な手法である PCA(Principal Component Analysis) を基準とし, それに対して, 構造保存, 疎性, 教師情報, 非線形性といった特定の性質を付加した手法との比較を通じて, 各設計思想の有効性を評価する.

線形変換としては,PCA,Random Projection,Sparse PCA, 及び Supervised PCA を採用する. まず,PCA はデータの分散を最大限保持する直交基底を用いる標準的な手法であり, 本研究の比較基準とする. 次に,Random Projection は計算効率の向上を重視した手法である. これは特徴量をランダムに低次元空間へ射影し, 学習過程を持たないため計算コストが極めて低い. また, 高確率で点対間距離が保存されることが理論的に保証されており [13], データの構造保持の有効性を検証する対象とする.

さらに, 特定の制約や情報を付与した線形変換として,Sparse PCA 及び Supervised PCA を導入する.Sparse PCA は, 主成分ベクトルに疎性の制約を課す手法であり, 解釈性の向上とともに不要な特徴量の影響を排除する効果がある. これを用いることで, 次元削減における疎性の考慮の有効性を評価する. 一方,Supervised PCA は, 目的変数との相関が高い特徴量を優先的に抽出する手法であり, 予測に直結する教師情報を活用することの有効性を検証する. なお, 実装については Bair らが提案した相関スクリーニングに基づく手法 [14] を採用した.

非線形変換としては,Transformer に基づく次元削減を採用する. 通常,Transformer は自己注意機構を用いて特徴間の複雑な非線形依存関係を表現できるが, 計算量が特徴量数の二乗に比例して増加するため, 高次元空間を対象とする本研究では計算コストが課題となる. そこで本研究では, 自己注意機構を線形化した Linear Transformer を用いることで, 非線形な表現能力と計算効率の両立を図る. これにより, 線形手法に対する非線形性の導入が Shapley 値の近似に与える影響を明らかにする.

パラメータにおいては, 各手法における次元削減の度合いが結果に与える影響を分析するため, 使用率を変更して実験を行う. 各次元削減手法について, 各データセットの特徴量次元に対して 5%,10%,20%の 3 通りの使用率を設定し, 使用率の違いによる計算効率及び説明誤差の変化を比較する.

ii-b) データセット

各次元削減手法の有効性が, データセットのどのような性質に依存するのかを詳細に分析するため, 特性の異なる複数のデータセットを用いる. 本研究では, 標準的なベンチマークである mnist [17] を基準とし,OpenML で広く使用されているデータセットである gisette [18],madelon [19],isolet [20] を組み合わせることで, 次元数, 相関構造, 疎密性の 3 つの観点から評価の軸を設定する. ここで, 各データセットの情報を表 1 にまとめる.

表 1　データセット

ID	データセット	次元数	相関構造	疎密
554	mnist	784	有	密
41021	gisette	5000	有	密
1485	madelon	500	無	密
300	isolet	617	有	疎

まず, 次元数の影響を検証するため,mnist に対して, 同様に相関構造を持ち密なデータセットである一方, より高次元な特徴量空間を持つ gisette を比較対象とする. これにより, 次元数の増大が各手法の計算効率および説明誤差に与える影響を明らかにする.

次に, 相関構造の影響を検証するため, 人工的に生成されたデータセットである madelon を採用し,mnist との比較を行う.madelon は, 目的変数とは無関係な特徴量や独立した分布からなるノイズを多く含み, データ全体としての有意な相関構造を持たない. この対比によって, 特徴間の相関構造の有無が与える影響を明らかにする.

さらに, 疎密性の影響を検証するため,mnist に対して isolet を検証対象とする.mnist が密なデータ構造を持つのに対し,isolet は音声信号を由来とする疎な構造を有している. これらの疎密性の違いに着目することで, データの疎密性が次元削減後の Shapley 値近似に及ぼす影響を分析する.

以上の 4 つのデータセットを通じて, 各次元削減手法が有効に機能する条件及びデータ特性との関係を明らかにする.

iii) 比較対象

次元削減を導入しない通常の Kernel SHAP をベースライン手法とする. また, 各次元削減を前処理として導入した Kernel SHAP を比較手法とする.

以上の設定に基づき, 計算時間,Shapley 値の値誤差, 及び特徴重要度ランキングの観点から性能を評価する.

4.3 実験結果

各データセットにおける実験結果の評価指標の数値をまとめた表を付録 B に記載する.

実験の結果, いずれのデータセットにおいても PCA,Random Projection, 及び Supervised PCA が次元削減未導入の場合よりも高い計算効率を達成した. 特に mnist においては, 次元削減未導入の場合と同程度の NormL1 を維持しており, 非常に良好な結果が得られた. 一方,Sparse PCA は値誤差では比較的良好な傾向を示したものの, 計算効率は他手法よりも極めて低い結果となった. また, 非線形変換である Linear Transformer については, 値誤差が極めて大きく, 計算効率の著しい低下が見られ

た. これらの傾向は, 次元数の高い gisette で特に顕著であった.

特徴重要度ランキングの観点では,mnist や gisette において,PCA,Supervised PCA 及び Sparse PCA が次元削減未導入の場合と同程度の Overlap@50 を示した. 特に Supervised PCA においては, 次元削減未導入の場合よりも高い計算効率で同程度の特徴構造を保持できている. また, 情報使用率を低く設定した手法ほど特徴構造を保持するという挙動も確認された.

4.4 次元削減手法の設計思想による分析

本節では, 各次元削減手法ごとの実験結果を抜粋して計算効率と説明誤差の関係を散布図として表し, 次元削減手法の設計思想が与える影響を分析する. なお, 図下部に記載の通り, 次元削減手法をプロットの色で区別している. また, 各図 (左) の両軸及び各図 (右) の横軸は対数スケールで表現している.

4.4.1 PCA / Random Projection：ランダム射影の影響

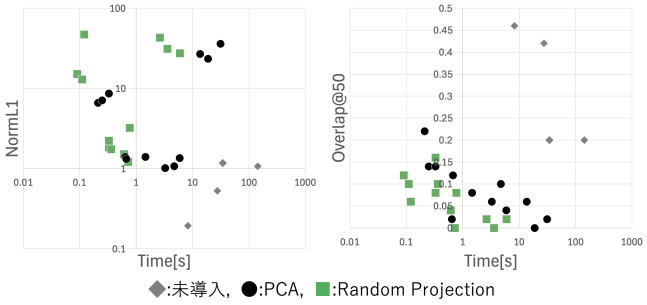

図 2　計算効率と説明誤差の関係 (PCA / Random Projection)

PCA は多くのデータセットにおいて計算効率と説明誤差の良好なトレードオフを実現しているのに対し,Random Projection は計算効率の面では PCA を上回るものの, 説明誤差においては PCA に劣る傾向が確認された. 特に情報使用率が低い条件において,Random Projection は PCA と比較して顕著に高い説明誤差を示した.

Random Projection は高次元空間の距離を低次元空間においても一定の誤差範囲内で保持することが理論的に保証されている [13]. しかし, 本実験の結果は,Shapley 値の近似には単なる点間距離の保持だけでは不十分であることを示唆している.PCA がデータの分散を最大化するように射影方向を決定して特徴構造を保持するのに対し,Random Projection はデータ構造を考慮せずに射影を行う.Shapley 値の算出には特徴量間の依存関係や相互作用の寄与が重要となるため, データ構造を無視したランダムな圧縮では, 近似に必要な情報が十分に低次元空間へ保持されなかったと考えられる.

以上の結果から, 特徴量間の複雑な相互作用に基づく Shapley 値を近似する場合, 計算コストの削減を優先してランダムな空間圧縮を行うよりも, 計算時間を一定程度許容してでもデータ固有の構造を直接的に反映する手法を選択することが, 説明誤差の抑制において不可欠であると結論付けられる.

4.4.2 PCA / Sparse PCA：疎性考慮の影響

Sparse PCA は, 値誤差では PCA と同程度であったものの,

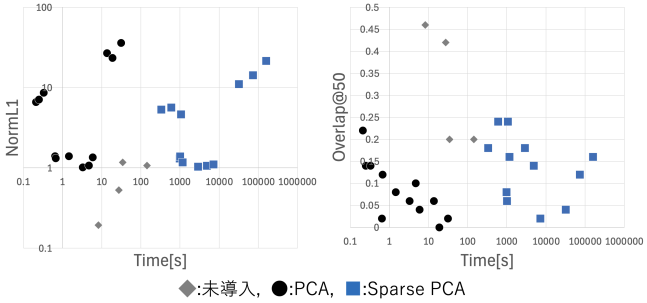

図 3　計算効率と説明誤差の関係 (PCA / Sparse PCA)

重要特徴の順位構造の保持においては,PCA と同等か, いくつかデータセットではそれを上回る良好な結果を示した. これは, スパース性を明示的に考慮することで,Shapley 値の算出に重要である寄与の大きい特徴量の構造がより明確に抽出され, ノイズの影響を排除した再構成が可能になったためと考えられる.

しかし, 計算効率の観点では,Sparse PCA は全データセットを通じて PCA を大きく上回る膨大な計算時間を要した. これは,Sparse PCA は L1 制約を伴う主成分係数の最適化問題を内部で反復的に解く必要があり, 特に高次元データにおいては反復計算のコストが爆発的に増大するためだと推測される [21].

以上より,Shapley 値の近似に次元削減を導入する場合, 特徴量の疎性を考慮することは重要特徴の順位構造の保持に一定の恩恵をもたらすものの, その計算コストの増大は構造保持という利点を大きく上回ることを示している. 実用的な観点からは, 同等の品質を遥かに短時間で達成できる PCA や後述する Supervised PCA の方が合理的な選択肢であると結論付けられる.

4.4.3 PCA / Supervised PCA：教師情報の影響

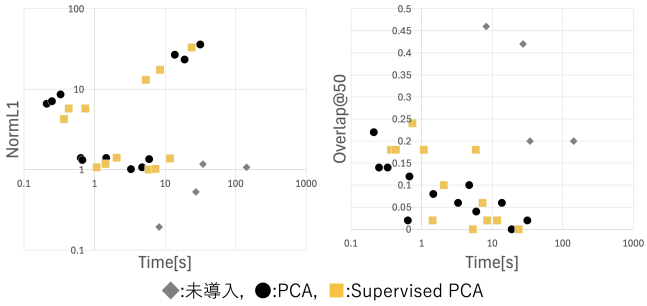

図 4　計算効率と説明誤差の関係 (PCA / Supervised PCA)

Supervised PCA は全てのデータセットにおいて PCA を上回る説明誤差を維持しつつ, 計算効率においても優秀な結果となった.

この結果は,Supervised PCA の設計思想が要因であると考えられる.Shapley 値は目的変数の予測に対する各特徴量の寄与度を算出する指標であるため, 次元削減の段階で目的変数の説明に寄与しない特徴量を教師情報を用いてあらかじめ排除する Supervised PCA の設計思想は,Shapley 値の定義に合致している. 通常の PCA が目的変数とは無関係な, 単に分散が大きいだけのノイズを保持してしまうリスクがあるのに対し,Supervised PCA は予測に直結する情報を選択して空間圧縮できるため, 説

明誤差を最小限に抑えられたと推測される.

以上の結果から, 単なるデータの分散に基づくよりも, 予測タスクの教師情報を次元削減プロセスに組み込むことが, 計算コストの削減と Shapley 値の説明誤差の抑制を両立させるための強力なアプローチとなると捉えることができる.

4.4.4 PCA / Linear Transformer：非線形性の影響

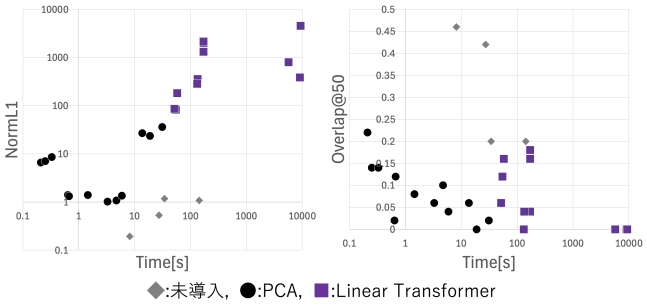

図 5 計算効率と説明誤差の関係 (PCA / Linear Transformer)

Linear Transformer は全てのデータセットにおいて, 他の手法と比較して極めて大きな説明誤差を記録した. また, 計算時間においても,Sparse PCA と同程度の高い計算コストを要した.

この深刻な説明誤差の爆発の要因としては,Shapley 値の数学的性質と,Transformer による非線形な埋め込み空間への写像との不整合が挙げられる.Shapley 値は加法性を有しており, 元の特徴空間における各要素の寄与を分解して評価する. しかし,Linear Transformer によって特徴量が複雑かつ非線形に混じり合った潜在表現へ変換されると, 元の特徴量と予測値の間の線形的な対応関係が破壊され, 正確な寄与度の再構成が困難になったと考えられる.

したがって,Shapley 値の再構成という目的においては,Transformer のような高度な非線形表現能力はむしろ過剰であり, 特徴空間の構造を直接的かつ明確に維持できる線形的な射影手法の方が, 計算コストと説明誤差の両面で圧倒的に適していると言える.

4.5 データセット特性による分析

本節では, 各データセットごとの実験結果を抜粋して計算効率と説明誤差の関係を散布図として表し, データセット特性による次元削減の影響の変化を分析する. なお, 図下部に記載の通り, データセットをプロットの色で区別している. また, 各図 (左) の両軸及び各図 (右) の横軸は対数スケールで表現している.

4.5.1 mnist / gisette：次元数の影響

mnist においては,PCA や Supervised PCA といった線形手法を用いることで, 次元削減を導入しない場合と比較して計算時間を大幅に短縮しつつ, 値誤差及び重要特徴の順位構造の両方において次元削減未導入とおおよそ同程度の性能を維持することが可能であった. これは, 中規模な次元数においては, 適切な射影基底を選択することで,Shapley 値の算出に必要な情報の大部分を低次元空間へ集約できることを示唆している.

一方で,5,000 次元を有する gisette においては, 次元削減による情報損失の影響がより大きく現れる傾向が確認された.PCA

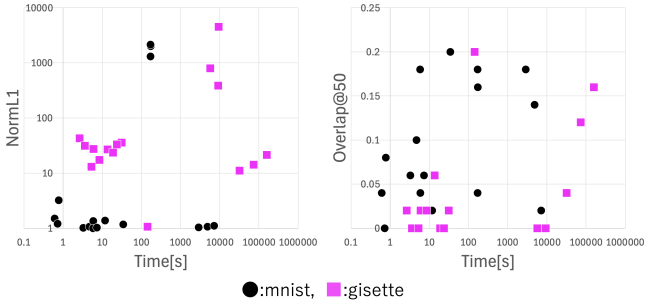

図 6 計算効率と説明誤差の関係 (mnist / gisette)

や Random Projection では,mnist と比較して Overlap@50 が著しく低下しており, 高次元空間においては次元削減に伴う僅かな情報損失が Shapley 値の順位構造を大きく損なう要因となる可能性がある.

4.5.2 mnist / madelon：相関構造の影響

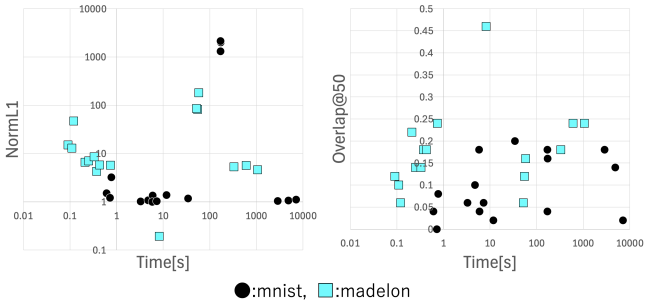

図 7 計算効率と説明誤差の関係 (mnist / madelon)

実験結果の全体的な傾向として, 評価指標によってデータセット特性の影響が異なることが確認された. 値誤差においては,mnist が madelon と比較して小さく良好な値を示した一方, 特徴重要度ランキングにおいては,madelon が mnist を上回る結果となった.

この結果は, データの相関構造が次元削減後の情報保持に与える影響が, 評価の側面によって二極化することを示唆している.mnist のように特徴量間に相関が存在する場合, データの本質的な情報が少数の成分に集約されやすいため, 個々の Shapley 値の絶対的な大きさの再現には有利に働く. しかし, 寄与の近い特徴量が密集しているため, 次元削減に伴う微細なノイズによって順位の逆転が生じやすく, ランキングの再現性は低下すると考えられる. これに対し,madelon は特徴量間の相関が乏しく個々の値誤差は増大しやすいものの, 予測に寄与する重要な特徴と無意味なノイズとの差異が明確であるため, 次元削減後も上位特徴の識別という観点では高い堅牢性を発揮したと解釈できる.

以上の結果から, 次元削減を用いた Shapley 値の近似においては, 説明の数値を重視するか, 重要特徴の特定を重視するかという目的に応じ, データの相関構造がもたらす恩恵と課題が表裏一体の関係にあるという特性が明らかとなった.

4.5.3 mnist / isolet：疎密性の影響

mnist と isolet の両データセット間において, 次元削減の導入に伴う説明誤差の推移や各手法の性能に, 決定的な差異は確

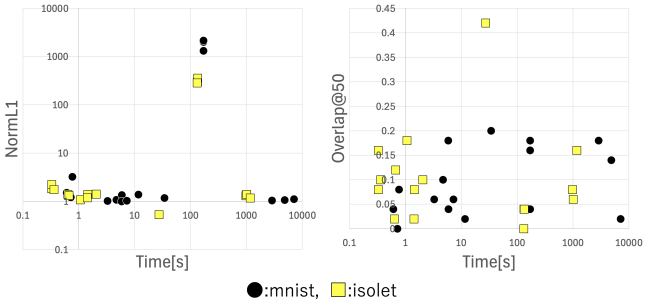

図 8　計算効率と説明誤差の関係 (mnist / isolet)

認されなかった. いずれのデータセットにおいても, 次元削減を適用することで一定程度の情報損失が生じるものの, その度合いや重要特徴の順位構造の保持傾向は概ね共通していた.

この結果は, 本研究で対象とした次元削減手法が, データの疎密性という性質に対して比較的堅牢であり, データの疎密性に関わらず設計思想に基づいた一定の有効性を発揮できることを示唆している. すなわち,Shapley 値の計算効率及び説明誤差を左右する主要な要因は, データの疎密性よりも, 前述した次元数や相関構造といった他のデータセット特性, あるいは次元削減手法自体の設計思想に強く依存するものと考えられる.

4.6　実用的な指針

本節では, 前節までの分析結果に基づき, モデル非依存 SHAP 近似の前処理として次元削減を適用する際の実用的な指針を, 次元削減手法の選択とデータ特性に基づく適用の判断の 2 つの観点からまとめる.

4.6.1　推奨される次元削減手法

次元削減を適用する場合, 計算効率と説明誤差のトレードオフが最も優れている Supervised PCA を第一の選択肢として推奨する.Supervised PCA は, 教師情報を活用することで予測に寄与する特徴量を効果的に抽出し, 計算コストを大幅に削減しつつ, ベースラインである PCA と比較して安定した説明品質を確保できる.

また, 説明誤差の許容範囲が広く, リアルタイム処理に近い高速性が求められる場合は,PCA や Random Projection が選択肢となる. ただし,Random Projection はデータ構造を無視するため, 説明誤差が大きくなるリスクを考慮する必要がある.

一方,Linear Transformer 等の非線形変換を用いる手法は,Shapley 値の加法性を損ない説明誤差を著しく増大させるため, 採用すべきではない. さらに,Sparse PCA は次元削減を適用しない場合よりも計算時間が大幅に増加し, 計算コストの削減という次元削減の本来の意義を損なうため, 実用性に乏しい.

4.6.2　データセット特性に基づく適用の判断

次元削減手法の選択以前に, 対象とするデータセットの特性に応じて, そもそも次元削減を適用すべきかを慎重に判断する必要がある. 本研究の実験結果は, 以下の傾向を示唆している.

まず,gisette の結果が示す通り, 次元数が極めて大きい場合, 次元削減に伴う情報損失の影響が顕著となり, いずれの手法を用いても説明誤差が増大する傾向にある. したがって, 極めて高い説明品質が要求されるタスクにおいては, 次元削減の適用を止めるか, 許容可能な誤差範囲を事前に厳密に定義する必要がある.

また, データの相関構造の有無は, 値誤差と順位構造に対してトレードオフの関係をもたらす. 相関構造を持つデータの場合, 次元削減によって情報の集約が効率的に行われるため, 値誤差は小さく抑えられる. しかし, 類似した特徴量が多数存在するため, 次元削減の微小なノイズによって順位が変動しやすい. したがって, 全体的な寄与の大きさを把握したい場合には次元削減は有効であるが, 厳密な重要特徴のランキングを特定したい場合には適さない可能性がある. これに対し, 相関構造が希薄なデータの場合, 値誤差は大きくなる傾向にあるが, 重要な特徴とノイズの区別が明確であるため, 順位構造は比較的保たれやすい. したがって, 重要特徴の特定を重視する場合には, 次元削減の適用が有効に機能する可能性がある.

以上のように, 単に計算時間の短縮のみを目的に次元削減を導入するのではなく, 対象データの次元数や相関構造, そして必要な説明特性を総合的に考慮して, 適用の可否を決定すべきである.

5　ま　と　め

本論文では, 高次元データにおける Shapley 値算出の計算コストの高さに着目し, 次元削減がモデル非依存型 SHAP 近似手法に与える影響を計算効率及び説明誤差の観点から体系的に評価した. 実験では, 性質の異なる 4 つのデータセット及び設計思想の異なる 5 つの次元削減手法を用い, 多角的な視点から比較検証した. その結果,PCA や Random Projection は説明誤差の観点で課題が残る一方, 教師情報を活用する Supervised PCA は特定の条件下では次元削減未導入時と同等の誤差抑制を達成しながら, 計算時間を大幅に短縮し, 全手法中で最善のトレードオフを示した. 対照的に,Sparse PCA は計算コストが爆発しており,Linear Transformer 等の非線形手法は Shapley 値の構造を損なうため, 不向きであることが明らかとなった.

今後の課題としては, 次元削減後の特徴空間において, より直接的に Shapley 値の構造を保持するような新しい次元削減アルゴリズムの開発が挙げられる. また, 本研究で得られた知見について, 異なるモデルやデータセットを用いた追加検証を通じて一般性を裏付けることで, 高次元データにおける XAI の実用性をさらに高めていく必要がある.

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである.

文　　献

[1] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin: ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", KDD2016
[2] Hsuan-Lei SHAO, Wei-Hsin WANG, Sieh-Chuen HUANG, Kuan-Ling SHEN: "Open the Black Box of AI:Saliency Map

of DUI Sentencing and Legal XAI", The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023
[3] DARPA, "XAI: Explainable Artificial Intelligence", https://www.darpa.mil/research/programs/explainable-artificial-in-telligence, 2025/02/12
[4] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, Evan W Newell: "Dimensionality reduction for visualizing single-cell data using UMAP", Nature Biotechnology, 2018/12/03
[5] Jiaxin Gao, Wenbo Hu, Yuntian Chen: "Revisiting PCA for time series reduction in temporal dimension", ICLR 2025, 2024/12
[6] Scott Lundberg, Su-In Lee: "A Unified Approach to Interpreting Model Predictions", NeurIPS 2017 Pages 4768-4777, 2017/12/04
[7] Scott M. Lundberg, Gabriel G. Erion, Su-In Lee: "Consistent Individualized Feature Attribution for Tree Ensembles", arXiv:1802.03888, 2018/02/12
[8] Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, David Camacho: "Exploring Dimensionality Reduction Techniques in Multilingual Transformers", Cognitive Computation Volume 15 pages 590–612, 2022/10/29
[9] Rabeeh Karimi Mahabadi, James Henderson, Sebastian Ruder: "Compacter: Efficient Low-Rank Hypercomplex Adapter Layers", NeurIPS 2021 Article No.79 Pages 1022-103, 2021/12/06
[10] Xuran Hu, Mingzhe Zhu, Zhengpeng Feng, Ljubiša Stanković: "Manifold-based Shapley explanations for high dimensional correlated features", Neural Networks Volume 180 106634, 2024/12
[11] Montgomery Flora, Corey Potvin, Amy McGovern, Shawn Handler: "Comparing Explanation Methods for Traditional Machine Learning Models Part 2: Quantifying Model Explainability Faithfulness and Improvements with Dimensionality Reduction", arXiv:2211.10378, 2022/11/18
[12] Stan Lipovetsky, Michael Conklin: “Analysis of regression in game theory approach”, Applied Stochastic Models in Business and Industry 17.4, pp. 319–330
[13] William B. Johnson, Joram Lindenstrauss: "EXTENSIONS OF LIPSCHITZ MAPPINGS INTO A HILBERT SPACE", Contemporary Mathematics Volume 26, 1984
[14] Eric Bair, Trevor Hastie, Debashis Paul, Robert Tibshirani: "Prediction by Supervised Principal Components", Journal of the American Statistical Association, Volume 101 Issue 473, 2006
[15] Hasso Plattner Institut, 「Introduction To Data Mining & Probabilistic Reasoning」, https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/folien/WS1213/IR/Intro4RetrievalEvaluation.pdf, 2026/01/03 参照
[16] BellCurve, 「スピアマンの順位相関係数」, https://bellcurve.jp/statistics/glossary/2052.html, 2026/01/03 参照
[17] OpenML, 「mnist」, https://www.openml.org/search?type=data&status=active&id=554, 2026/01/25 参照
[18] OpenML, 「gisette」, https://www.openml.org/search?type=data&status=active&id=41026, 2026/01/02 参照
[19] OpenML, 「madelon」, https://www.openml.org/search?type=data&status=active&id=1485, 2026/01/02 参照
[20] OpenML, 「isolet」, https://www.openml.org/search?type=data&status=active&id=300, 2026/01/25 参照
[21] scikit-learn, 「SparsePCA」, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.SparsePCA.html, 2026/01/10 参照

付 録 A

本実験は以下の計算機環境において実施した.

ハードウェア環境

デバイス：MacBook Air (M1, 2020)
チップ：Apple M1 (8 コア CPU, 8 コア GPU)
メモリ：8GB

ソフトウェア環境

OS：macOS Sonoma 14.4.1
Python：3.11.7

使用ライブラリとツール

XGBoost：2.1.3
SHAP：0.46.0
scikit-learn：1.2.2
NumPy：1.26.4
SciPy：1.11.4
PyTorch：2.2.2

付 録 B

データセット mnist,gisette,madelon,isolet による実験結果の各評価指標の具体的な数値をそれぞれ表 2,3,4,5 に示す.

表 2　実験結果 (mnist)

手法名(使用率)	NormL1	RMAPE	Overlap@10	Overlap@50	Spearman	Time[s]
未導入	1.17	1.33	0.70	0.20	0.23	34.29
PCA(5%)	1.01	1.15	0.00	0.06	-0.03	3.28
PCA(10%)	1.07	1.21	0.00	0.10	-0.03	4.73
PCA(20%)	1.35	1.53	0.00	0.04	0.00	5.92
Random Projection(5%)	3.22	3.65	0.00	0.08	-0.01	0.77
Random Projection(10%)	1.51	1.71	0.00	0.04	0.03	0.61
Random Projection(20%)	1.21	1.38	0.00	0.00	0.06	0.72
Sparse PCA(5%)	1.04	1.18	0.60	0.18	0.00	2873.79
Sparse PCA(10%)	1.07	1.21	0.10	0.14	0.01	4872.06
Sparse PCA(20%)	1.11	1.26	0.00	0.02	0.01	7124.47
Supervised PCA(5%)	1.00	1.14	0.00	0.18	0.09	5.84
Supervised PCA(10%)	1.02	1.16	0.00	0.06	0.08	7.28
Supervised PCA(20%)	1.38	1.56	0.00	0.02	-0.04	11.73
Linear Transformer(5%)	1997.74	2269.68	0.20	0.16	0.02	171.86
Linear Transformer(10%)	1313.55	1492.36	0.00	0.04	-0.03	170.56
Linear Transformer(20%)	2138.15	2429.20	0.00	0.18	-0.04	170.72

表 3　実験結果 (gisette)

手法名(使用率)	NormL1	RMAPE	Overlap@10	Overlap@50	Spearman	Time[s]
未導入	1.07	0.13	0.60	0.20	0.15	142.68
PCA(5%)	26.81	3.31	0.00	0.06	0.04	13.65
PCA(10%)	23.47	2.90	0.00	0.00	0.02	18.87
PCA(20%)	35.90	4.44	0.00	0.02	0.02	31.33
Random Projection(5%)	42.92	5.31	0.00	0.02	-0.01	2.64
Random Projection(10%)	31.26	3.86	0.00	0.00	-0.04	3.58
Random Projection(20%)	27.35	3.38	0.00	0.02	-0.03	5.94
Sparse PCA(5%)	11.07	1.37	0.10	0.04	0.01	32183.80
Sparse PCA(10%)	14.24	1.76	0.10	0.12	0.01	72786.07
Sparse PCA(20%)	21.48	2.65	0.10	0.16	0.00	158422.34
Supervised PCA(5%)	13.05	1.61	0.00	0.00	0.05	5.32
Supervised PCA(10%)	17.42	2.15	0.00	0.02	0.03	8.47
Supervised PCA(20%)	33.08	4.09	0.00	0.00	-0.01	23.77
Linear Transformer(5%)	385.07	47.60	0.00	0.00	0.03	9127.80
Linear Transformer(10%)	795.40	98.32	0.00	0.00	0.03	5677.98
Linear Transformer(20%)	4526.41	559.50	0.00	0.00	0.01	9357.35

表 4　実験結果 (madelon)

手法名(使用率)	NormL1	RMAPE	Overlap@10	Overlap@50	Spearman	Time[s]
未導入	0.19	0.13	0.80	0.46	0.51	8.22
PCA(5%)	6.62	4.45	0.20	0.22	-0.12	0.21
PCA(10%)	7.07	4.75	0.40	0.14	-0.05	0.25
PCA(20%)	8.61	5.78	0.60	0.14	-0.08	0.33
Random Projection(5%)	47.03	31.60	0.00	0.06	-0.07	0.12
Random Projection(10%)	15.06	10.12	0.00	0.12	-0.04	0.09
Random Projection(20%)	12.90	8.67	0.10	0.10	0.01	0.11
Sparse PCA(5%)	5.31	3.57	0.50	0.18	0.05	331.09
Sparse PCA(10%)	5.65	3.80	0.50	0.24	0.08	600.98
Sparse PCA(20%)	4.63	3.11	0.50	0.24	-0.01	1058.32
Supervised PCA(5%)	4.25	2.86	0.40	0.18	0.03	0.37
Supervised PCA(10%)	5.77	3.88	0.20	0.18	-0.03	0.43
Supervised PCA(20%)	5.75	3.86	0.40	0.24	-0.07	0.74
Linear Transformer(5%)	182.04	122.34	0.00	0.16	0.05	57.92
Linear Transformer(10%)	82.16	55.21	0.00	0.12	-0.02	54.71
Linear Transformer(20%)	85.55	57.49	0.10	0.06	0.07	51.91

表 5　実験結果 (isolet)

手法名(使用率)	NormL1	RMAPE	Overlap@10	Overlap@50	Spearman	Time[s]
未導入	0.53	0.37	0.70	0.42	0.59	27.37
PCA(5%)	1.40	0.98	0.00	0.02	0.04	0.64
PCA(10%)	1.32	0.92	0.00	0.12	0.02	0.67
PCA(20%)	1.40	0.98	0.00	0.08	0.00	1.46
Random Projection(5%)	1.83	1.28	0.10	0.08	0.04	0.33
Random Projection(10%)	2.22	1.55	0.00	0.16	-0.04	0.33
Random Projection(20%)	1.75	1.22	0.00	0.10	0.02	0.36
Sparse PCA(5%)	1.31	0.91	0.00	0.08	-0.10	982.41
Sparse PCA(10%)	1.39	0.97	0.00	0.06	-0.04	1006.51
Sparse PCA(20%)	1.17	0.82	0.10	0.16	0.02	1168.92
Supervised PCA(5%)	1.07	0.75	0.10	0.18	-0.01	1.07
Supervised PCA(10%)	1.18	0.82	0.00	0.02	-0.09	1.43
Supervised PCA(20%)	1.41	0.98	0.10	0.10	0.01	2.06
Linear Transformer(5%)	295.57	206.29	0.00	0.04	-0.01	133.48
Linear Transformer(10%)	354.72	247.57	0.00	0.04	-0.03	134.59
Linear Transformer(20%)	284.34	198.45	0.00	0.00	-0.04	131.72

LLM の Unknown-Unknown を捉える Human-in-the-Loop エンティティマッチング

岡山　紘汰†　　伊藤　寛祥††　　森嶋　厚行††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305-0821 茨城県つくば市春日 12
†† 筑波大学 図書館情報メディア系　〒 305-0821 茨城県つくば市春日 1-2
E-mail: †kota.okayama.2025b@gmail.com, ††{ito,morishima-office}@slis.tsukuba.ac.jp

あらまし　エンティティマッチングは，データベースに存在する同一の実体を統合する作業として知られている．大規模言語モデル（LLM）はエンティティマッチングにおいて高い性能を発揮する一方，モデルが自信を持って誤った判断を下す「Unknown-Unknown」問題が依然として課題である．本研究は，この問題が推移律の矛盾として表面化することに着目する．この矛盾を手がかりにアノテーション用のデータを選別する矛盾駆動型の能動学習戦略 LLM に適応し，その有効性を検証する．具体的には，3 つのエンティティ間で推移律が破綻する「矛盾した三角形」を検出し，その矛盾度合いをスコア化する．そして，矛盾スコアが高いペアを優先的にアノテーションの対象として選別し，得られたラベル付きデータをモデルへフィードバックを行った．日本語の公立図書館の書誌データ，英語の音楽・人物・商品データなど，特性の異なる複数のデータセットを用いて提案手法の評価を行い，比較手法と比べ精度が向上することを示した．

キーワード　エンティティマッチング，大規模言語モデル（LLM），能動学習，推移律，Human-in-the-Loop

1　序　　論

エンティティマッチング（Entity Matching: EM）は，異なるデータソース間に散在する実世界の同一エンティティを参照するレコードを特定・統合するための操作であり，データ品質管理における基本的かつ重要なプロセスである [1]．実世界のデータは，表記ゆれ，欠損，あるいはスキーマの不一致といったデータを含んでおり，従来はルールベースや機械学習を用いた手法が研究されてきた．近年では，大規模言語モデル（Large Language Models: LLM）がエンティティマッチングタスクにおいて顕著な性能を示しており，その豊富な事前知識により，意味的なマッチングが可能となっている．

しかし，機械学習モデルや LLM の活用には重大な課題が残されている．それは，モデルが高い確信度（Confidence）を持ちながら誤った予測を行う「Unknown-Unknown 問題」である [2]．LLM も従来のモデルと同様に，判断が難しい事例に対しては確信度が低くなる傾向にあるが，誤った判断に対しても極めて高い確信度を示す場合がある．不確実性サンプリング（Uncertainty Sampling）[3] のような従来の能動学習戦略は，モデルが判断に迷う（不確実な）事例を検出することはできるが，こうした「Unknown-Unknown」な誤りを特定することは困難である．この課題を明確にするため，我々は「Unknown-Unknown」状態を構成する 2 つの要素を定義する．すなわち，Unknown (1) は根本的な「誤り」つまり，実際の正解とは異なる判断をしたことであり，Unknown (2) はその誤りの検出を阻む「不確実性の欠如（高い確信度スコア）」である．本研究は，この隠れた誤り群をターゲットとする．

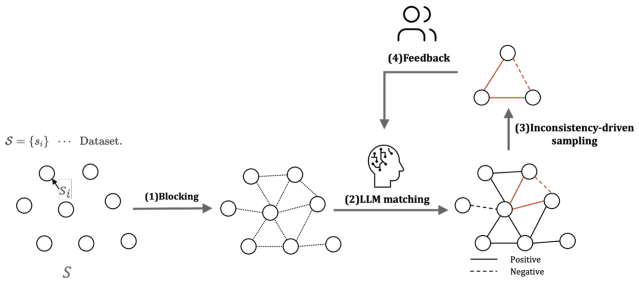

図 1　矛盾駆動型 Human-in-the-loop LLM エンティティマッチング

推移律の矛盾（Inconsistency）を利用して EM における「Unknown-Unknowns」に対処するアプローチが存在する [4]．この手法は，高確信度の誤りがしばしば論理的な矛盾，特に推移律（すなわち，$A = B$ かつ $B = C$ ならば $A = C$）の違反として現れるという洞察に基づいている．これらの矛盾なペアに焦点を当てることで，彼らはマッチング精度の効率的な向上を達成した．しかし，彼らの検証は古典的な機械学習モデルに限定されていた．そのため，この戦略の現代的な LLM への適用は体系的に調査されておらず，例えばファインチューニング（FT）と Few-shot プロンプティングのどちらが適切かなど，矛盾から得られた情報をどのようにフィードバックするのが最も効果的かは不明なままである．

したがって本研究のリサーチクエスチョン（RQ）は以下の通りである．

- RQ1: 矛盾駆動型アプローチは，LLM ベースのエンティティマッチングにおいてどの程度有効か？
- RQ2: Few-shot と FT では，どちらのフィードバック戦略

がより効果的か，またそれはどのような条件やデータセットにおいてか？

本研究においては,「矛盾駆動型（Inconsistency-driven）」サンプリング戦略が, 限られた予算の下で, 特に「Unknown-Unknown」問題に焦点を当てつつ，EM における LLM の精度を効果的に向上させることができるかを体系的に検証する．図 1 に本戦略の概要を示す．プロセスは入力されたエンティティ集合に対して次のように進行する．(1) ブロッキングにより，マッチする可能性の高い候補ペア（ノード）を絞り込む．(2) LLM が各候補ペアの一致確率を計算し，これらの確率を重みとするエッジを持つ無向グラフを構築する．(3) グラフ内の各三角形について，推移律違反の度合いを表す矛盾スコアを計算する．最もスコアが高い三角形から得られるユニークなペア集合を，人間の対象として選出する．(4) 最後に，人間が検証したラベルを LLM にフィードバックする．

本研究では，我々の手法を詳述するとともに，同一のアノテーション予算の下で複数のサンプリング戦略（矛盾駆動型，不確実性，ActiveLLM [5]，ランダム）およびフィードバックメカニズム（Few-shot とファインチューニング（FT））を比較した実験結果を提示する．

これらを踏まえ，具体的には，書誌，音楽，人物，および商品データセットを用いた実験を行い，F1 スコアの差を評価した．その結果，RQ1 に関して，実験結果は，LLM 単独で達成可能な最大性能はデータセットやタスク設定によって異なるものの，矛盾駆動型の選択戦略は多くのケースにおいて一貫して改善をもたらすことを示している．RQ2 に関しては，多くのデータセットにおいて FT（具体的には矛盾由来のサンプルを用いたファインチューニング）が特に効果的であり，Few-shot プロンプティングも特定の条件下で有益であることがわかった．これらの知見は，EM における LLM の矛盾という課題に取り組む上で，推移律に基づく価値の高いサンプルのアノテーションを優先する戦略が非常に有効であることを示している．

2 関 連 研 究

本章では，関連研究を紹介し，本研究の立ち位置を明確にする．

2.1 古典的機械学習および LLM を用いた EM 手法

既存のエンティティマッチング（EM）手法には，文字列距離，トークンマッチング，あるいは人間が作成したルールやその合成に基づく古典的アプローチが含まれる [6], [7], [8]．また，特徴量エンジニアリングとランダムフォレスト等の分類器を組み合わせた機械学習手法 [9], [10] や，クラウドソーシングを通じた人間の判断を取り入れる Human-in-the-Loop システム，およびメトリック学習に基づく類似度学習 [11], [12] も存在する．近年では，大規模言語モデル（LLM）をマッチャーとして用いる手法が登場し，EM タスクに対する強力なアプローチとして確立されている [13].

2.2 マッチングモデルの性能改善手法

特定のタスクに対して LLM の性能を自動的に向上させる取り組みも行われている．例えば，Ji らによる Autonomous Learning は，人間が参考書を用いて学習するプロセスを模倣するものである [14]．改善プロセスへの人間の関与を前提とする本研究とは異なり，この手法はモデルが自身に欠落していると認識できない情報からは学習できない可能性がある．また，別の能動学習手法である ActiveLLM [5] は，学習データの選別に LLM を使用するが，これも LLM 自身の「Unknown-Unknowns（確信を持った誤り）」を捉えられないという課題を抱えている．

プロンプトエンジニアリングの観点からは，Wang らが EM における 3 つの戦略，すなわち Matching（ペアワイズ判定），Comparing（2 候補間の比較），Selecting（リストからの選択）を比較し，性能向上を調査した [15]．改善の方向性は多岐にわたり，外部知識を統合する RAG（Retrieval-Augmented Generation）[16] や，EM に特化したファインチューニング [17] なども含まれる．

2.3 推移律を用いたエンティティマッチングの改善手法

推移律を利用して，マニュアルでエンティティマッチングの推論を補助・改善することで EM 性能を高める試みもある．Zhu らによる推論補助と整合性に関する研究は，クラウドソーシング環境を前提とし，推移律に基づいて人間が手動でモデルの回答に対して修正を行うものである [18]．本研究は，推移律を「モデルの回答における推移率の矛盾の検出とアノテーションを行うデータの選別」の指標として能動学習に組み込み，アノテーションを行ったデータの情報価値を高める点でこれとは異なる．

推移律を能動学習のためにエンティティマッチングモデルに適用した例として，Ito らはベイズ分類器における「Unknown-Unknown」問題，すなわち高い確信度を持ちながら誤った予測を行う問題に対処する手法を提案した [4]．彼らの手法では，3 つのエンティティ間のマッチング結果が推移律に違反する「矛盾」を検出し，矛盾の原因となっているペアを提示して人間の介入を促す．これにより，不確実性サンプリングのような従来の能動学習手法では見過ごされがちな，モデルが確信を持って誤っている事例を効率的に発見することが可能となる．本研究とは，LLM を主要なマッチャーに据えた点，それにより，モデルに対するフィードバック手法がパラメータチューニングのみである点が異なる．

2.4 本研究の位置づけ

本研究の新規性は，LLM を主要なマッチャーとして位置づけた上で，推移律由来の矛盾スコアに基づいてアノテーションペアを選別することでエンティティマッチングにおける LLM の性能が向上するかを検証する点にある．さらに，このフィードバックをファインチューニングや fewshot プロンプトなど LLM に対してどのように与えるのが最も効果的かを探求する．

3 問題設定

本稿では，エンティティマッチング（EM）を，2 つのレコードが実世界における同一の存在（エンティティ）を指しているか否かを判定する 2 値分類問題として定式化する．理解を容易にするため，まず具体的な入力例を示す．図 2 は，音楽データセット [19] に含まれる 2 つのレコードペアの例である．各レコード r は，$\{\text{Title}, \text{Length}, \text{Artist}, \text{Album}, \text{Year}\}$ といった属性集合を持ち，これらの属性値のテキスト表現が LLM への入力となる．

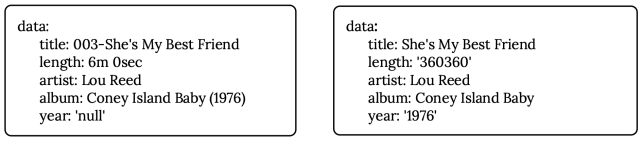

図 2　音楽データセットにおけるレコードペアの入力例

我々は，LLM を確率的マッチャー f として採用する．ペア x が入力された際，マッチャー f はマッチングの予測ラベル $\hat{y} \in \{\text{“Match”}, \text{“Unmatch”}\}$ に加え，その予測に対する確信度（信頼度スコア）を出力する．本研究では，この確信度をマッチング確率 $P(\text{Match}|x)$ と解釈し，閾値を 0.5 として以下のように判定を行う．

- $P(\text{Match}|x) \geqq 0.5$ の場合: 同一エンティティ（Match）と予測．
- $P(\text{Match}|x) < 0.5$ の場合: 異なるエンティティ（Unmatch）と予測．

具体的な LLM の出力フォーマットは以下の通りである．

Example Output: Yes. Match Probability: 0.95

問題をより一般化して定義する．エンティティの集合を S とし，マッチング対象となるすべての可能なレコードペアの集合を $X = \{(s_i, s_j) \mid s_i, s_j \in S, i < j\}$ とする．各ペア $x \in X$ には，真のラベル $y \in Y = \{0, 1\}$ が存在する（ここで 1 は Match，0 は Unmatch に対応する）．正解ラベルが付与された全データの集合を $\mathcal{U}_n \subset X \times Y$ と定義し，そのサイズを $n = |\mathcal{U}_n|$ とする．

能動学習（Active Learning）の枠組みにおいて，我々は全データ $\mathcal{U}_n$ の中から，予算制約 m （$m \ll n$）を満たす小さな部分集合 $L_m \subset \mathcal{U}_n$ を選択し，人間によるアノテーションを取得することができる．ここで，マッチャー $f : X \to Y$ は，学習データ（あるいは Few-shot プロンプト）に基づいて仮説空間 $\mathcal{F}$ から選択される関数である．

本研究の目的は，限られたアノテーション予算 m 内で，システム全体の正解率を最大化する最適な選択戦略（クエリ関数）$Q : \mathcal{U}_n \to L_m$ を設計することである [4]．これは，以下の目的関数を最大化する部分集合 L_m を探索する問題に帰着される．

$$\underset{L_m \subseteq \mathcal{U}_n, |L_m| = m}{\operatorname{argmax}} \frac{1}{n} \sum_{(x,y) \in \mathcal{U}_n} \delta\left(f(x) = y \vee (x, y) \in L_m\right) \quad (1)$$

ここで $\delta(\cdot)$ は条件が満たされた場合に 1 を返す指示関数である．この式 (1) の総和の中身は，**「モデル f が正解する」**または**「人間がアノテーションを行い正解ラベルを与えている $((x, y) \in L_m)$」**のいずれかの場合に，そのデータは正解であるとみなすことを意味する．すなわち，単純にモデルの予測精度を上げるだけでなく，モデルが苦手とする（誤分類する可能性が高い）事例を L_m としてサンプリングし，その数を最大化，人間が正解を与えることで，モデルにとって Unknown-Unknown なデータを学習させ，性能向上を図るのが本研究の狙いである．

4 提案手法

4.1 フレームワークの概要

本節では，図 1 に示す矛盾駆動型能動学習の全体フレームワークについて詳述する．ワークフローは以下の通り進行する．

1. **ブロッキング**: マッチする可能性が最も高い候補ペアの集合に絞り込むためにブロッキングを適用する．
2. **推論と矛盾検出**: 現在のモデルを使用して全候補ペアのマッチ確率を推論し，そこから推移律に違反する「矛盾した三角形」を計算・抽出する．
3. **サンプリング**: これらの矛盾した三角形を利用し，矛盾度の高いペアを優先的に人間のアノテーション対象とする．
4. **フィードバック**: 新たにアノテーションされたペアをモデルにフィードバックする．

4.2 ブロッキング

エンティティマッチングの計算コストを削減するため，まずブロッキング処理を適用して非マッチングペアを除去し，候補ペアの集合を効率的に絞り込む．ブロッキングは本研究の主眼ではないため，汎用的な埋め込み表現を用いた近傍探索に基づく標準的なブロッキング手法を採用する．手順をアルゴリズム 1 に示す．

Algorithm 1 Nearest Neighbor Blocking

Input: $S = \{s_1, \ldots, s_N\}$: A set of entities; $E : S \to \mathbb{R}^d$: An embedding function; k: The number of neighbors
Output: $\mathcal{X}_{\text{cand}}$: A set of candidate pairs

$V \leftarrow \{E(s) \mid s \in S\}$ ▷ Embed all entities into a vector space
index $\leftarrow$ BuildSearchIndex(V)
$\mathcal{X}_{\text{cand}} \leftarrow \emptyset$
for each entity $s_i \in S$ **do**
    $N_i \leftarrow$ index.search($E(s_i), k$) ▷ Find k nearest neighbors for s_i
    for each neighbor $s_j \in N_i$ **do**
        $\mathcal{X}_{\text{cand}} \leftarrow \mathcal{X}_{\text{cand}} \cup \{(s_i, s_j)\}$
    end for
end for
return $\mathcal{X}_{\text{cand}}$

まず，アルゴリズムは全エンティティ集合 $S = \{s_1, \ldots, s_N\}$，埋め込み関数 E，および各エンティティに対して探索する近傍数 k を入力とする．アルゴリズムは最初に，埋め込み関数 E

を用いてすべてのエンティティ $s \in S$ を d 次元のベクトル空間に変換し，ベクトル集合 V を生成する（行 1）．次に，効率的な近傍探索を可能にするために，これらのベクトルからインデックスを構築する（行 2）．

インデックス構築後，各エンティティの埋め込みベクトル $E(s_i)$ に対して k 個の最近傍エンティティ s_j を探索する（行 4）．本研究では，埋め込みベクトル間の L2 距離を使用した．元のエンティティ s_i と発見された近傍 s_j からなるペア (s_i, s_j) が候補ペア集合 $\mathcal{X}_{\text{cand}}$ に追加される（行 6）．このプロセスを全エンティティに対して繰り返すことで，マッチする可能性が高いペアのみを含む精錬された候補ペア集合 $\mathcal{X}_{\text{cand}}$ が得られる．

4.3 矛盾駆動型サンプリング

図 1 のパート (3) に対応する矛盾駆動型サンプリングアルゴリズムをアルゴリズム 2 に示す．

Algorithm 2 Inconsistency-driven sampling

Input: $\mathcal{X}_{\text{cand}}$: A set of candidate pairs; f: An LLM-based matcher; m: Annotation budget
Output: Q: A set of newly labeled pairs

1: $P \leftarrow \{(x, f(x)) \mid x \in \mathcal{X}_{\text{cand}}\}$ ▷ Predict match probabilities
2: $T \leftarrow \text{FindTriangles}(\mathcal{X}_{\text{cand}})$
3: $T_{\text{scores}} \leftarrow \emptyset$
4: **for** each triangle $t \in T$ **do**
5: $\quad \text{score} \leftarrow \text{CalculateInconsistency}(t, P)$
6: $\quad T_{\text{scores}} \leftarrow T_{\text{scores}} \cup \{(t, \text{score})\}$
7: **end for**
8: $T_{\text{sorted}} \leftarrow \text{SortByScore}(T_{\text{scores}})$ ▷ Sort triangles by inconsistency score
9: $Q \leftarrow \text{GetUniquePairsFromTopTriangles}(T_{\text{sorted}}, m)$
10: **return** Q

このアルゴリズムは，候補ペア集合 $\mathcal{X}_{\text{cand}}$，LLM マッチャー f，およびアノテーション予算 m を入力とする．まず，初期モデル f を使用して各候補ペア $x \in \mathcal{X}_{\text{cand}}$ のマッチ確率を予測し，結果を P として保存する．

次に，アルゴリズムの核心部分である矛盾計算に移る．候補ペア集合内の 3 つのエンティティ $\{s_a, s_b, s_c\}$ によって形成されるすべての三角形構造 $t = \{(s_a, s_b), (s_b, s_c), (s_c, s_a)\}$ を特定する（行 2）．

図 3 は，本サンプリング戦略がターゲットとする論理的に矛盾が生じている 3 つのエンティティを示している．この図において，実線はマッチしたペア（Match），点線は非マッチのペア（Unmatch）を表す．矛盾する三角形（Contradictory triangle）とは，モデルが 2 つのマッチと 1 つの非マッチ（例：Match-Match-Unmatch）を予測し，推移律に違反しているエンティティ群として明示的に定義される．特定された各三角形 t について，その 3 つのエッジに対応する予測確率 $p(a, b), p(b, c), p(c, a)$ を使用して「矛盾スコア」を計算する（行 4-7）．このスコアは，3 つのエンティティ内の論理的矛盾の度合いを定量化するために特別に設計されており，図 3 に視覚的に示されるような矛盾した三角形を検出する．

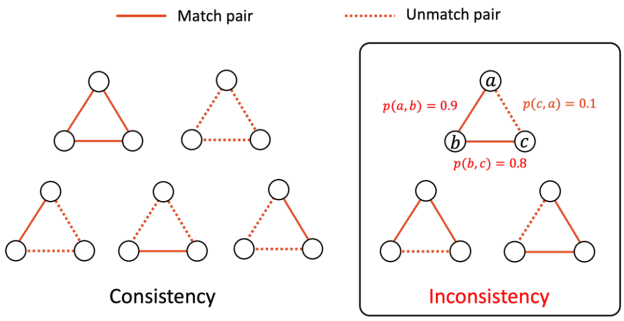

図 3 矛盾したデータの例

このスコアは以下の式 (2) で定義される．

$$\text{score} = \sum_{(i,j,k) \in \text{cyc}(a,b,c)} p(i, j)p(j, k)(1 - p(k, i)) \qquad (2)$$

ここで，$\text{cyc}(a, b, c)$ は (a, b, c) およびその巡回置換 $(b, c, a), (c, a, b)$ の集合を表す．

すべての三角形の矛盾スコアを計算した後，アルゴリズムはそれらを降順にソートする．その後，アノテーション予算 m に達するまで，最も矛盾な三角形からペアを選択してクエリ集合 Q を生成する（行 8-9）．これらのペアは，現在のモデルが不確実または誤っている可能性が高い「最も情報量の多い」サンプルと見なされる．

最後に，選択されたクエリ集合 Q は人間に提示されて正解ラベルを取得し，ラベル付き集合 L を作成する（行 10）．このプロセスは，限られたアノテーション予算の制約内でモデルのパフォーマンスを最大化することを目的としている．

4.4 フィードバック戦略 (Few-shot / FT)

本研究のフレームワークは，選択 → アノテーション → モデル更新 → 再推論のサイクルに従う．本研究では，アノテーションされたデータ L_m をモデルにフィードバックするための 2 つの主要な戦略，すなわち Few-shot 学習とファインチューニング（FT）を比較する．

両戦略に共通する点は，矛盾スコアの高いペアが優先的にアノテーションされ，所定数のアノテーション済み事例が「学習データ」として使用されることである．各戦略におけるサンプルサイズやハイパーパラメータの具体的な設定については次章で詳述する．

Few-shot 戦略では，この学習データからいくつかの事例（例：矛盾スコアが最も高いペア）が選択され，推論時のコンテキスト情報としてプロンプトに埋め込まれる．これにより，モデルの重みを変更することなく，特定のタスクに対するモデルの応答を調整する．

対照的に，**ファインチューニング（FT）**は，学習データを使用してベース LLM の重みを直接更新し，タスクに特化したモデルを生成する．

どちらの戦略においても，公平な性能評価を保証するために，フィードバックに使用される学習データと性能測定に使用され

る評価データは，相互に排他的な（共通部分を持たない）集合として扱われる．

5 実　　験

本節では，RQ1 および RQ2 に答えるために実施された実験とその結果について述べる．評価は，多様なドメインをカバーする 5 つのデータセットを用いて行われた．

5.1 データセット

Persons [19] は，英語の個人属性からなる比較的クリーンなデータセットであり，ベースラインのマッチング性能を評価するために使用される．**Bibliorecords** は，日本の公立図書館から提供された日本語の書誌データセットであり，表記の揺れに対するモデルの堅牢性を測定するために使用される．**Music** [19] は，多くの欠損値やノイズ属性を含む英語の音楽データセットであり，ノイズの多い実データを扱う能力を評価するために使用される．

さらに，製品ドメインから 2 つのデータセットを選択した．**Amazon-Walmart** [20] は，異なる EC サイトの製品データを対象としている．**WDC-Products** [21] は，ウェブアーカイブから大規模に収集された実際の製品データに基づくベンチマークである．このベンチマークは異なる指標に基づく複数のデータセットで構成されているが，本研究ではデータの 80% がコーナーケース（判断が難しい事例）で構成されているバージョン「WDC-Products 80%」を使用する [21]．

表 1 に各データセットの概要と実験で使用した属性を示す．

表 1　データセットの属性

ドメイン (言語)	データセット	属性
人物 (英語)	Persons	名, 姓, 市区町村, 郵便番号
書誌 (日本語)	Bibliorecords	タイトル, 著者, 出版社, 日付
音楽 (英語)	Music	アーティスト, タイトル, アルバム, 年, 長さ
製品 (英語)	Amazon-Walmart	タイトル, 価格, ブランド, モデル番号
製品 (英語)	WDC-Products 80%	ブランド, タイトル, 説明, 価格

実験では，これらの各データセットからサンプリングされた約 3,000 レコードのサブセットを使用した．各データセットを学習用 2,000 レコード，評価用 1,000 レコードに分割した．プロンプトで LLM に入力する際には，表 1 の属性を 1 つの文字列に連結した．これらの属性に加えて，データセット内のすべてのレコードには「クラスタ ID」という特別な属性がある．同じデータセット内のレコードペアが同じクラスタ ID を持つ場合，実験の評価フェーズではそれらのレコードがマッチしていると見なす．

表 2 は，本実験のために構築された評価用サブセットの統計情報を示している．本研究では，提案手法の汎用性を多角的に評価するため，人物（Persons），書誌情報（Bibliorecords），音楽（Music），および E コマース商品（Amazon-Walmart, WDC-Products）という，ドメインの異なる 5 つのデータセットを採用した．

特筆すべきは，これらのデータセットがドメインの違いだけでなく，データの構造的特性においても顕著な差異を持っている点である．例えば，`Bibliorecords` は平均クラスタサイズが 3.43 と大きく，同一エンティティが多く含まれる「密 (Dense)」なデータセットである．対照的に，`Amazon-Walmart` は 1000 レコードに対しマッチ数がわずか 38 件しか存在せず，極めて不均衡かつ疎な性質を持っている．このように，マッチング頻度やクラスタ構造が大きく異なるデータセット群を用いることで，特定のデータ分布に依存せず，多様な難易度や条件下において提案手法が堅牢に機能するかを体系的に検証することが可能となる．

表 2　実験用データセットの統計

データセット	レコード数	クラスタ数	平均サイズ	マッチ数
Persons	1000	637	1.57	1570
Bibliorecords	1001	292	3.43	1817
Music	1002	491	2.04	888
Amazon-Walmart	1000	963	1.04	38
WDC-Products	1000	500	2.00	500

5.2 実 験 設 定

5.2.1 実験ワークフロー

マッチャー，データセット，予算 m を伴うサンプリング戦略，およびフィードバック戦略が与えられたとき，実験ワークフローは以下のように進行する．

1. **ブロッキング:** データセットからのデータ項目のペア集合に対し，アルゴリズム 1 を適用する．ここでは，埋め込みベースの近傍探索を採用する．具体的には，`text-embedding-ada-002` から得られた埋め込みベクトルを使用し，FAISS [22] の IndexFlatL2 を利用して，ユークリッド距離（L2）に基づき各レコードの上位 10 件の候補（$k = 10$）を特定する．
2. **初期マッチャー f_0 の性能測定:** 上記の手法でブロッキングを行った後，抽出されたペア集合に対し f_0 を適用した際の F1 スコアを計算する．
3. **サンプリング戦略の適用:** 抽出されたペアへのマッチャーの実行結果に対して，予算 m でサンプリング戦略を適用し，m 個のペアを選択する．
4. **モデル更新:** フィードバック戦略 p を用いてモデル f_0 を更新し，f_1 を得る．
5. **更新後マッチャー f_1 の性能測定．**

5.2.2 モデルアーキテクチャと学習設定

本研究では，ベースモデルとして OpenAI の `gpt-4o-mini-2024-07-18` を採用した．同モデルは，高い推論能力を維持しつつ，コスト効率と応答速度に優れており，大規模な実験を行う本研究のフレームワークに適している．

a） 学習データの構築

モデルのファインチューニング（SFT: Supervised Fine-Tuning）には，OpenAI Fine-tuning API を使用した．サンプリング戦略によって選択されたレコードペア（L_m）は，チャット補完（Chat Completion）形式の JSONL フォーマットに変換され，学習データとしてモデルに供給される．具体的には，各トレーニング事例は以下の 3 つの役割（Role）によって構成される．

- **System:** モデルの役割定義．ここでは「音楽エンティティのマッチングを行う専門家」として振る舞うよう指示を与え，出力形式（Yes/No および確信度スコア）を規定する．
- **User:** 入力データ．マッチング対象となる 2 つの音楽レコードの属性情報（タイトル，アーティスト，アルバム等）を提示する．
- **Assistant:** 期待される出力（正解ラベル）．アノテーション結果に基づき，マッチする場合は「Yes」，しない場合は「No」とともに，1.0 または 0.0 のスコアを規定する．

b） ハイパーパラメータ設定

学習時のハイパーパラメータとして，バッチサイズは 1，学習率（Learning Rate Multiplier）は 1.8，エポック数は 3 に設定した．また，推論時の生成におけるランダム性を排除し，結果の再現性を確保するため，Temperature パラメータは 0 に設定している．

5.3 ベースラインと比較手法

実験では，以下のように LLM へのフィードバックを行った．

- **ファインチューニング (FT):** 前述の推移律不整合を通じて検出されたペアを使用して，LLM のパラメータを直接更新する手法．
- **Few-shot:** LLM のプロンプト内に少数の正解例を提示する手法．今回の実験では推移律違反を起こしている三角形 4 つ分のデータを Few-shot 事例として組み込む．

提案する不整合駆動型サンプリングの有効性を検証するため，以下のベースラインおよび比較手法を設定した．

- **Zero-shot:** 能動学習を行わない，事前学習済み LLM の初期性能．
- **ランダムサンプリング (Random):** 全候補ペアからランダムにデータを選択し，それを LLM にファインチューニングして比較する．
- **不確実性サンプリング (Uncertainty) [3]:** モデルが予測に最も迷っている（LLM の確信度が低い）データを優先する手法．具体的には $|P(s_i, s_j) - 0.5|$ が最小となるレコードペアを選択する．
- **Selecting 戦略 [15]:** 1 つのアンカーレコードと 10 個の候補レコードをプロンプト内で提示し，一致する候補を選択させる．
- **ActiveLLM [5]:** LLM 自身に次にすべきデータを選択させる手法．200 件の候補から LLM が 32 件を選択するプロセスを繰り返す．

公平な比較を行うため，各サンプリング戦略で選択されるユニークなペアの数は，不整合駆動型サンプリングで選択される数と一致させている．

5.4 評価設定

主な指標は，クラスタから導出されたペアワイズ F1 スコアである．予算は「100 個の不整合三角形に含まれるユニークなペアの数」に基づいており，この数はすべての比較手法で揃えられている．

6 結果と考察

6.1 主要な性能評価

提案する矛盾駆動型サンプリングの有効性を評価するため，複数のデータセット（Bib, Music, Person, Amazon-Walmart, WDC-product）において性能を比較した．100 個の矛盾三角形をサンプリングした後の F1 スコアを表 3 に示す．

表 3 実験結果：異なるサンプリング戦略における F1 値の比較

戦略	Persons	Bib	Music	Amz-Wal	WDC
Zero-shot	0.8849	0.9894	0.9680	0.8182	0.6103
Random	0.9803	0.9688	0.9833	0.7170	0.5421
Uncertainty	0.9955	**0.9932**	0.9855	**0.9459**	0.6299
ActiveLLM	0.9745	0.9929	0.9916	0.1242	0.1879
Selecting	0.9070	0.9677	0.9887	0.3684	0.4336
Ours (Few-shot)	0.9305	0.9826	0.9787	0.9067	0.6126
Ours (FT)	**1.0000**	**0.9932**	**0.9960**	0.8861	**0.6820**

表 3 は，各戦略で選択されたデータを用いてファインチューニング（FT）を行った後の性能を示している．実験結果から，提案手法は複数のデータタイプにおいて，既存手法（Zero-shot, Random, Uncertainty, ActiveLLM）と比較して同等以上の性能を一貫して達成していることが確認された．person データセットでは F1 スコアが 1 を達成し他の比較手法と比べても優れた値を示した，Bib データセットでは，提案手法を用いた FT が F1 スコア 0.9932 を達成し，最も性能の高かった比較手法である Uncertainty サンプリングと同等であった。Music データセットでは，提案手法は 0.9960 という F1 スコアを記録し，他のすべての手法を明確に上回った。同様に，WDC-product データセットにおいても，提案手法は最高の F1 スコア 0.6820 を記録した。一方，Amazon-Walmart データセットでは，Uncertainty 手法が最高の性能（0.9459）を示したのに対し，提案手法による性能向上は他のデータセットほど高くはなかった。この原因については次節で分析する。

6.2 矛盾三角形の量の分析

Amazon-Walmart での性能向上が限定的であった理由を調査するため，各データセットの矛盾三角形の量を分析した（表 4）．

表 4 学習データにおける矛盾三角形の数

データセット	全ペア数	全三角形数	矛盾三角形数
Persons	21,179	65,820	418
Bibliorecords	23,406	60,356	295
Music	22,450	15,505	304
Amazon-Walmart	14,024	24,032	20
WDC-Products	13,034	30,681	2,089

Amazon-Walmart には矛盾三角形が 20 個しか含まれていなかった．これはマッチするペアが極めて少ない「疎」なデータセットであるため，三角形構造自体が形成されにくいことに起因する．この結果から，提案手法の有効性は矛盾三角形の十分な存在に依存することがわかる．

本節では，実験結果に基づき，序論で提示したリサーチクエスチョンに回答する．

RQ1: 矛盾駆動型アプローチは，LLM ベースのエンティティマッチングにおいてどの程度有効か？実験結果は，矛盾駆動型サンプリングが広範に有効であることを実証している．提案手法は，4 つのドメイン（書誌，音楽，人物，WDC 製品）において，ベースラインと比較して優れた F1 スコアを達成した．特に Person データセットでは，F1 スコア 1.0 という完璧な性能に到達した．これは，推移律に違反するサンプルを優先することで，ランダムや標準的な不確実性サンプリングよりも効果的にモデル内部の「Unknown-Unknown」エラーを修正できることを経験的に証明している．

RQ2: Few-shot と FT では，どちらのフィードバック戦略がより効果的か？フィードバック戦略の比較に関して，分析の結果，ファインチューニング（FT）が一般的に効果的であるが，最適な戦略はデータセットの特性に依存することが明らかになった．大多数のデータセットにおいて，推移律矛盾由来のサンプルを用いた FT が高い性能をもたらし，エンティティマッチングタスクにおける堅牢性が確認された．しかし，重要な発見として，推移率矛盾を起こしている三角形が少ないデータセット（例：Walmart-Amazon）では例外が見られ，このようなケースでは，Few-shot が FT を上回った．これは，価値の高い矛盾事例が希少な場合，それらをプロンプトで直接提示する方が，モデルパラメータを更新するよりもデータ効率が良くモデルの性能が向上しやすいことを示していると考えられる．

7 限界と今後の課題

提案手法の有効性は示されたが，いくつかの限界と今後の課題が存在する．

第一に，矛盾サンプリングの有効性は，候補ペア間に推移律矛盾を起こした三角形が十分に存在することに依存している．データセットにおいて，エンティティ間の関係が疎（スパース）である場合，矛盾検出の機会が減少し，サンプリング手法の効果が限定的なものとなる可能性がある．多くの主流なエンティティマッチングデータセットは 1 対 1 のマッチングを中心に構成されているか，エンティティあたり数個のマッチングペアしか含まないため，我々の手法の利点が顕著に現れにくいシナリオも存在する．

第二に，我々の評価は商用 LLM API に依存しており，予告なしのモデル更新による性能変動などの外的要因の影響を受けやすい．しかし，能動学習フレームワーク自体は他の LLM にも適応可能である．

第三に，我々の結果は，推移律違反が主に「真であるべきペアが誤って偽と予測される（False Negative）」ケースから生じていることを示唆している．現在のサンプリング手法は，スコアの高い矛盾三角形内のすべてのペアを候補として検討するため，モデルが既に正しく分類しているペアを含んでしまう可能性があり，非効率である．これを解決する戦略として，矛盾を直接引き起こしているペア，具体的には推移律に違反する誤った予測（例：(A, B) と (B, C) が真と予測されているのに，ペア (A, C) を偽と予測している場合）を持つペアを優先的にサンプリングすることが考えられる．今後は，このピンポイントなサンプリング戦略を LLM 向けに実装し，さらに少ない予算でモデル性能を最大化できるかを検証する予定である．

8 結　　論

本研究では，推移律の矛盾に着目した，LLM ベースのエンティティマッチングのための新しい能動学習戦略を提案した．複数のデータセットを用いた評価実験の結果，本手法はランダムサンプリングや不確実性サンプリングなどの既存手法と比較して，安定して上回るか同等の性能を達成することが示された．特に，フィードバック手法としてファインチューニングと組み合わせた場合，モデル性能の向上における有効性が確認された．

これらの知見は，サンプリング基準として LLM の判断の論理的整合性を使用することの妥当性を裏付けており，将来の能動学習研究に対する新たな方向性を示唆している．今後の課題として，より多くのモデルへの適用による検証の拡大，より大規模で多様なデータセットでのテスト，および他の論理的制約を組み込んだサンプリング戦略の探求が挙げられる．

謝　　辞

本研究の一部は JSPS 科研費 (22H00508，23K28095，25K21807) と JST CREST(Grant Number JPMJCR22M2) の支援を受けたものである．ここに謝意を示す．

文　　献

[1] Nils Barlaug and Jon A Gulla. Neural networks for entity matching: A survey. *ACM Computing Surveys*, Vol. 15, No. 3, pp. 1–37, 2021.

[2] Yeounoh Chung, peter J Haas, Eli Upfal, and Tim Kraska. Unknown examples & machine learning model generalization, 2019.

[3] Burr Settles. Active learning literature survey, 2010. Technical Report, University of Wisconsin-Madison, `https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf`.

[4] Hiroyoshi Ito, Takahiro Koizumi, Ryuji Yoshimoto, Yukihiro Fukushima, Takashi Harada, and Atsuyuki Morishima. Inconsistency-driven approach for human-in-the-loop entity

matching. *Information Research an International Electronic Journal*, Vol. 30, No. iConf, pp. 1024–1038, 2025.

[5] Markus Bayer, Justin Lutz, and Christian Reuter. Activellm: Large language model-based active learning for textual few-shot scenarios. *arXiv preprint arXiv:2405.10808*, 2025.

[6] Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, Vol. 84, No. 406, pp. 414–420, 1989.

[7] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: a generic approach to entity resolution. *The VLDB Journal*, Vol. 18, No. 1, pp. 255–276, 2009.

[8] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, Vol. 3, pp. 73–78, 2003.

[9] Chaitanya Gokhale, Saravanan Das, Anhai Doan, Jeffrey F Naughton, Narasimhan Rampalli, Jude Shavlik, and Xiaojin Zhu. Corleone: Hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 601–612, 2014.

[10] Sanjib Das, Paul Suganthan G.C., AnHai Doan, Jeffrey F. Naughton, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, Vijay Raghavendra, and Youngchoon Park. Falcon: Scaling up hands-off crowdsourced entity matching to build cloud services. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1431–1446, 2017.

[11] Naofumi Osawa, Hiroyoshi Ito, Yukihiro Fukushima, Takashi Harada, and Atsuyuki Morishima. Bubble:a quality-aware human-in-the-loop entity matching framework. In *The 5th IEEE Workshop on Human-in-the-loop Methods and Future of Work in Big-Data (IEEE HMData2021)*, pp. 3557–3565, 2021.

[12] Harada Takashi, Fukushima Yukihiro, Sato Sho, Tsuruta Misato, Yoshimoto Ryuji, and Morishima Atsuyuki. Advancement of bibliographic identification using a crowdsourcing system. In *Proceedings of the 9th Asia-Pacific Conference on Library & Information Education and Practice (A-LIEP 2019)*, pp. 71–82, 1993.

[13] Ralph Peeters, Aaron Steiner, and Christian Bizer. Entity matching using large language models. *arXiv preprint arXiv:2310.11244*, 2023.

[14] Ke Ji, Junying Chen, Anningzhe Gao, Wenya Xie, Xiang Wan, and Benyou Wang. Unlocking llms' self-improvement capacity with autonomous learning for domain adaptation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21051–21067, 2025.

[15] Tianshu Wang, Xiaoyang Chen, Hongyu Lin, Xuanang Chen, Xianpei Han, Hao Wang, Zhenyu Zeng, and Le Sun. Match, compare, or select? an investigation of large language models for entity matching. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 96–109, 2025.

[16] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

[17] Aaron Steiner, Ralph Peeters, and Christian Bizer. Fine-tuning large language models for entity matching. *arXiv preprint arXiv:2409.08185*, 2025.

[18] Yao Zhu, Hongzhi Liu, Zhighai Wu, and Yingpeng Du. Relation-aware neighborhood matching model for entity alignment. *arXiv preprint arXiv:2012.08128*, 2020.

[19] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. In *Proceedings of the VLDB Endowment*, Vol. 3, pp. 484–493, 2010.

[20] Hasso Plattner Institute. Amazon-walmart product matching dataset, 2025. Retrieved from `https://hpi.de/naumann/projects/repeatability/datasets/amazon-walmart-dataset.html` (Accessed 2025-08-29).

[21] Ralph Peeters, Reng C Der, and Christian Bizer. Wdc products: A multi-dimensional entity matching benchmark. *arXiv preprint arXiv:2301.09521*, 2023.

[22] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, Vol. 7, No. 3, pp. 535–547, 2019.

複数エージェントの学習戦略を表現する解釈可能なルール集合の獲得

今村 優志† 太田 学†† 上野 史††

† 岡山大学工学部工学科情報・電気・数理データサイエンス系 〒 700–8530 岡山県岡山市北区津島中 3-1-1
†† 岡山大学学術研究院環境生命自然科学学域 〒 700–8530 岡山県岡山市北区津島中 3-1-1
E-mail: †pdiq2xao@s.okayama-u.ac.jp, ††{ohta, uwano}@okayama-u.ac.jp

あらまし 近年，説明可能な AI（eXplainable Artificial Intelligence: XAI）は，人とロボットの協働における必要技術として着目されている．中でも，正確性に基づく学習分類子システム（eXtended Classifier System: XCS）は，個々の状況に適合するルールの集合を知識として学習することで，ロボットのナビゲーションなどの問題において適切な行動戦略を人が解釈できるルールとして保持できる AI システムである．しかし，XCS はルールの正確さを基準としてルールを最適化するため，最適行動が他のエージェントに依存するマルチエージェント環境では適切なルール獲得が困難であるという問題がある．そこで，本研究では，他エージェントに対する競争的戦略と協調的戦略を示す二つのルール集合に基づき学習する XCS を提案し，マルチエージェント迷路環境における解釈可能なルール集合の獲得を目指す．実験結果より，この二つのルール集合による学習を行うことで，二体のエージェントが存在するマルチエージェント迷路環境において，解釈可能なルール集合を獲得できることが示された．また，単一のルール集合のみを用いる XCS と比較して，ゴール到達までのチームの平均ステップ数のより早い収束が確認された．

キーワード 説明可能な AI，強化学習，遺伝的アルゴリズム，学習分類子システム，マルチエージェント経路探索

1 はじめに

近年，人工知能（Artificial Intelligence: AI）技術の発展が著しく，現実社会における問題に適用される例が増えている．一方で，現在主流の AI は極めて複雑なモデルであり，人間にとって判断根拠の把握が困難なブラックボックスモデルである．特に，人間とロボットの協調的な行動が必要とされる場面において，ロボットの意思決定の根拠が人間にとって理解困難であることが問題視されている [1].

説明可能な AI（eXplainable Artificial Intelligence: XAI）は，AI による出力の根拠を抽出する技術であり，近年需要が高まっている．その中でも，学習分類子システム（Learning Classifier System: LCS）は解釈可能なルール集合を知識として表現し，強化学習（Reinforcement Learning）と遺伝的アルゴリズム（Genetic Algorithm: GA）によって環境に対する最適方策を学習するシステムである．LCS は環境との相互作用を通じて獲得したルールを一般化することで，汎用的に適用可能な知識の獲得を目指す．特に，正確性に基づく学習分類子システム（eXtended Classifier System: XCS）は，ルールの正確性（ルールを実行することで得られる報酬値の確からしさ）を指標にするため，AI によって解決する問題の構造理解に役立つシステムである．

XCS はこれまで単体のエージェントと静的な環境を対象としていた．しかし，現実環境は動的であり，特に複数ロボットの協調制御等に用いられるマルチエージェントシステムでは，自身の行動の最適性が他のエージェントの振る舞いに依存して変化するため，正確なルール集合の獲得が困難であるという問題がある．

以上の背景を踏まえ，本研究ではマルチエージェント環境における XCS を新たに提案する．具体的には，他のエージェントに対して競争的戦略と協調的戦略の二つのルール集合に基づき学習する XCS を提案し，他のエージェントの行動に応じた最適ルールを保持するように拡張する．本研究では，複数エージェントによる迷路問題を取り上げ，自己のゴール到達を中心に学習する競争的ルール集合と，自己と他エージェントが共にゴールへ到達することを目指し学習する協調的ルール集合の二つを持った XCS を提案する．

また本論文では，提案システムの有効性を検証するため，マルチエージェント迷路環境において実験を実施する．比較対象として，単一のルール集合のみを有する従来の XCS を用い，両手法を同一条件下で評価する．実験結果に基づき，学習回数に対する最適経路までの収束性能と，提案手法が獲得したルール集合の性能について評価する．

本論文の構成は以下のとおりである．まず，2 節では XCS や XCS で用いる迷路環境に関する関連研究，3 節では迷路探索問題について紹介する．次に，4 節では XCS のアルゴリズムについて述べ，5 節では提案手法，6 節では実験方法，実験結果，および考察を述べる．7 節ではまとめと今後の課題を述べる．

2 関連研究

2.1 学習分類子システム

学習分類子システム（Learning Classifier System: LCS）[2] は，J. H. Holland により提案された，強化学習と遺伝的アルゴリズムを組み合わせた進化的機械学習手法である．LCS は if-then ルールと付随するパラメータからなる分類子の集合を最適化することで環境に適応する．具体的には，環境からの観測

に適合する分類子をルール集合から選び出し，ルールに従った行動を選択し，環境からの報酬に基づき分類子のパラメータを更新することで，環境に適応したルール集合を獲得する．

しかし，LCS による分類子獲得は，得られた報酬の大きさに基づいて評価が行われるため，分類子の予測正確性が十分に考慮されていないという問題が存在する．その結果，予測誤差が大きいにも関わらず，高い報酬を偶発的に獲得した分類子が行動選択に強く影響を及ぼす可能性がある．

そこで，S. W. Wilson は正確性に基づく学習分類子システム（eXtended Classifier System: XCS）[3] を提案した．XCS は，LCS の特徴に加え，分類子の正確性，すなわちルール実行によって得られる報酬予測の信頼性に基づいてルールを最適化し，環境適応性の高い包括的なルール集合を形成する．XCS は，環境における各状態において行動空間を網羅的に学習することにより，分類問題や迷路探索問題のような未知環境において安定した環境適応能力を示してきた [3]．一方で，状態遷移の不確実性が想定されるマルチエージェント迷路環境においては，安定した学習が困難であると予想される．

LCS や XCS の知識獲得の大きな特徴として，複数の入力状態に照合する汎用的な分類子を獲得する一般化が挙げられる．一般化により，必要最小限の分類子集合で広い状態をカバーすることが可能となる．

2.2 迷路探索問題における XCS

迷路探索問題は，エージェントが未知の環境において，障害物などを回避しつつ目標へ到達するための経路を学習する問題であり，XCS における代表的なベンチマークの一つである．XCS による迷路探索問題では，Woods2 環境 [3]，Maze6 [4] などの Grid Worlds 型の迷路環境において実験が行われてきた．Woods2 環境において，XCS は最適行動を学習しつつ，獲得した知識を適切に一般化できることが示されてきた．一方で，Maze6 環境は障害物が多く複雑な環境であり，標準的な XCS の一般化メカニズムでは過剰な一般化が起こりやすく，最適経路の行動の学習が難しいとされている．

迷路探索問題において，従来は知覚エイリアシングと呼ばれる，異なる場所で同一のセンサ情報を観測することで起こるエージェントの誤認に対して，正確なルール集合の獲得を目指した研究が進められてきた．知覚エイリアシングに対する最も直感的なアプローチは，メモリを用いて観測情報の時系列から現在の状態を認識することである [5]．Lanzi は，XCS にメモリを導入した XCSM（XCS with Internal Memory）を提案し，単純な非マルコフ環境において最適解に収束し，内部メモリのサイズに対して安定性を示すことを明らかにした [6]．その一方で，予測に基づく XCS として Butz は Anticipatory Classifier System 2（ACS2）[7] を提案しており，現在では知覚エイリアシングに対して主流の手法となっている．また近年では，Siddique らが時系列情報から特徴的なパターンを抽出して利用する FoRsXCS（Frames-of-References-based XCS）[8] を提案し，上野らはそれを連続した知覚エイリアシングに対して改良した AGFX2（Aliasing-Group-based FoRsXCS with Dual-stream Identification）[9] を提案した．

	0	1	2	3	4	5	6	7	8
0	T	T	T	T	T	T	T	T	T
1	T						T	F	T
2	T			T		T	T		T
3	T		T						T
4	T				T	T			T
5	T		T		T			T	T
6	T		T						T
7	T						T		T
8	T	T	T	T	T	T	T	T	T

図 1　Maze6 の詳細図

以上の問題は，知覚エイリアシングの発生場所は変化しない．一方で，マルチエージェント環境では，他のエージェントの振る舞いによって誤認が発生する場所が動的に変化するため，従来とは質の異なる難しさを含んでいる．特に，動的な環境では正確なルール集合の獲得が困難であり，本研究では従来の研究の流れを踏襲しつつより現実的なシナリオに拡張している．

3 迷路探索問題

迷路探索問題は，XCS の学習性能および一般化性能を検証するための代表的なベンチマークとして用いられてきた．XCS による迷路探索問題の迷路環境は，Grid Worlds 型の迷路環境であり，各セルは通路（空セル）と，障害物としての木（T 記号セル），および食物（F 記号セル）から構成される．環境に配置されたアニマット（エージェント）は食物セルへの到達を目的として行動し，環境への適応を行う．アニマットは，隣接するセルのうち通路セルへ移動することが可能であるが，移動先が木セルである場合には移動できない．一方で，移動先セルが食物である場合，そのセルに移動し，食物獲得による報酬を得る．アニマットは周囲 8 近傍の環境状態を観測可能であり，各方向の状態を 2 ビットで符号化する．具体的には，通路を “00”，木を “01”，食物を “11” として表現する．よって 2 ビット ×8 方向の 16 ビット長のビット列を一つの状態表現として用いる．行動は観測方向に対応する 8 種類から構成され，上方向を 0 とし，時計回りに 0 から 7 の数値で表現される．

Maze6 の構造を図 1 に示す．セル内の文字が状態を表し，上辺と左辺の数字はそれぞれ x 座標，y 座標を示す．Maze6 の場合，例えば，食物（ゴール）は $(1,7)$ に配置されている．

Maze6 の構造を図 1 に示す．セル内の文字が状態を表し，上辺と左辺の数字はそれぞれ x 座標，y 座標を示す．Maze6 の場合，例えば，食物（ゴール）は $(x,y)=(7,1)$ に配置されている．

4 正確性に基づく学習分類子システム

4.1 分　類　子

XCS は，環境の一部に適合する条件-行動ルールの集合を用いて学習することで，目的達成のためのルールを解釈性を維持し

表 1　分類子の構成要素

要素名	説明
条件部 (condition: C)	環境のパターン
行動部 (action: A)	分類子の提案する行動
予測値 (prediction: p)	得られると予想される報酬値
誤差値 (error: ϵ)	予測報酬値と報酬値の誤差の平均
適合度 (fitness; F)	報酬の予測の正確さ
重複度 (action set size: as)	選択時の平均行動集合サイズ
経験値 (experience: exp)	分類子が選択された回数
重合度 (numerosity: num)	同一分類子の集約数

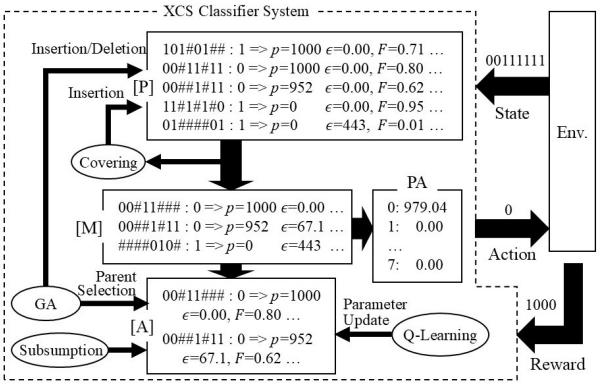

図 2　XCS の処理の概要

つつ最適化することができる．具体的には，表 1 に示す，条件-行動ルールとそれに付随するパラメータを分類子（Classifier）として持つ．分類子の条件部は，基本的に “0”, “1” に加え，これらのどちらも受け付ける “#”(ドントケア) の符号列で表現される．“#” により，XCS の一般化が実現される．

4.2　XCS の処理の流れ

XCS の処理の流れを図 2 に示す．[P] は XCS が所有する分類子集合，[M] は観測した状態と一致する条件部を持つ分類子の集合，PA は各行動の予測報酬，[A] は [M] 内にある選択された行動を提案した分類子の集合を示している．XCS は，次の六つの処理を指定された回数繰り返し実行し，環境に最適なルールの取得を試みる．

1. 環境の観測による状態の取得
2. 状態に適合する条件部を持つ分類子の選択
3. 各行動の予測報酬の計算
4. 行動の選択と実施
5. Q-Learning によるパラメータ更新
6. 遺伝的アルゴリズムによる分類子の生成と削除

図 2 を用いて，XCS のメカニズムについて説明する．まず，XCS エージェントは，環境（Env.）から状態（State）を観測し，分類子集合（Population: [P]）の中から，状態と適合する条件部を持つ分類子を抽出し，新たに照合集合（Match set: [M]）を構成する．なお，[M] が空集合，あるいは条件を満たす分類子が不足している場合は新たな分類子を生成するカバリング処理（Covering）が行われる．カバリングについては 4.5.1 項で述べる．

そして XCS は，[M] に含まれる分類子の予測値（p）と適合度（F）に基づき，実行可能な各行動に対する予測報酬 $P_t(a)$ を計算し，予測配列（PA）を生成する．予測報酬は式 (1) で計算される．

$$P_t(a_i) = \sum_{cl_k \in [M](a_i)} cl_k.p \times \frac{cl_k.F}{\sum_{cl_j \in [M](a_i)} cl_j.F} \tag{1}$$

なお，$P_t(a_i)$ は t 番目のステップにおける行動 a_i の予測報酬，$cl_j \in$[M](a_i) は [M] 内で行動 a_i を提案する分類子，$cl_j.p$，$cl_j.F$ それぞれは cl_j の予測値と適合度である．

その後，XCS は PA の値に基づき行動（Action）を決定する．同時に，[M] の中からその行動を行動部に持つ分類子のみを抽出し，行動集合（Action set: [A]）という分類子の集合を形成する．選ばれた行動を実施し，エージェントは環境から報酬を受け取り，XCS は分類子のパラメータを更新する．

4.3　強化学習によるパラメータ更新

XCS は，報酬に基づいて [A] 内の各分類子のパラメータを更新する．この更新は，Q-learning [13] に基づき実行される．Q-learning は，状態における行動の行動価値関数を学習することで，予測報酬が最大となる行動方策を獲得する強化学習手法である．価値の更新には TD 誤差（Temporal Difference Error）が用いられ，現在の価値予測と，将来の報酬を考慮した価値予測との差に基づいて行動価値が修正される．XCS では，この TD 誤差に基づき，各分類子のパラメータが更新される．具体的には，表 1 における予測値（p），誤差値（ϵ），適合度（F）を更新する．以下にこれらの更新式とその説明を述べる．なお $s, a,$ はそれぞれ状態，行動を表している．

予測値は式 (2) で更新される．

$$p_j \leftarrow p_j + \beta[r + \gamma \max_a P(s', a) - p_j] \tag{2}$$

ここで，$p_j, r, s', \max_a P(s', a), \beta, \gamma$ はそれぞれ予測値，報酬，次状態，次状態における最大の予測報酬，学習率，割引率を表している．右辺括弧内は TD 誤差に相当し，実際の経験（報酬と次状態の価値）と現在の予測値 p_j との差を表している．この更新式により，各分類子の予測値は，更新を繰り返すことで環境の真の報酬値へと収束する．

次に誤差値の更新について，誤差値は式 (3) で更新される．

$$\epsilon_j \leftarrow \epsilon_j + \beta[|r + \gamma \max_a P(s', a) - p_j| - \epsilon_j] \tag{3}$$

右辺の絶対値項は今回の行動実行による予測のずれの大きさを表している．この式は，過去の絶対誤差の平均値を推定するものであり，ϵ_j が 0 に近いほど，その分類子は環境の報酬を正確に予測できていることを意味する．

誤差値の更新後，分類子の正確さを計算する．XCS において適合度は、環境から得られる報酬の絶対量ではなく、予測誤差の小ささ、すなわち予測の正確性に基づいて決定される。まず、予測誤差 ϵ_j を用いて、絶対的な正確さ κ_j を式 (4) により計算する．

$$\kappa_j = \begin{cases} 1 & \text{if } \epsilon_j \leqq \epsilon_0 \\ \alpha(\epsilon_j/\epsilon_0)^{-\nu} & \text{otherwise} \end{cases} \tag{4}$$

ここで，ϵ_0 は許容誤差閾値，α および ν は κ_j を計算するためのパラメータである．誤差 ϵ_j が閾値 ϵ_0 以下であれば，その分類子の絶対的な正確さは 1 であるとみなされる．一方で，誤差が大きい場合は指数的に絶対的な正確さは低下する．

次に，計算した κ_j を用いて，[A] 内での相対的な正確さ κ'_j を式 (5) で計算する．

$$\kappa'_j = \frac{(\kappa_j \times num_j)}{\sum_{cl_k \in [A]} (cl_k.\kappa \times cl_k.num)} \tag{5}$$

ここで，num は表 1 の重合度 (Numerosity) を表す．この式により，その分類子の [A] 内の他の分類子との相対的な正確さの比率が得られる．

最後に κ'_j を用いて適合度を式 (6) により更新する．

$$F_j \leftarrow F_j + \beta[\kappa'_j - F_j] \tag{6}$$

この更新則により，XCS は，単に高い報酬を得るルールではなく，報酬値を正確に予測できるルールの適合度が高くなる．

4.4 遺伝的アルゴリズムによる分類子の生成と削除

XCS では，分類子の生成と削除に遺伝的アルゴリズム (Genetic Algorithm: GA) [10] を用いる．GA は，環境への適応度に基づいて分類子を選択し，交叉および突然変異といった遺伝的操作を通じて新たな分類子を生成する手法である．また，適合度の低い分類子の削除も GA により行われる．XCS では、環境からの入力に対して行動集合 [A] が構築された際，分類子が生成されてからの平均時間を確認する．この時間が閾値 θ_{GA} を超えた場合，遺伝的アルゴリズムが発動する．XCS における遺伝的アルゴリズムでは，親の選択，交叉，突然変異を実施する．それに加え，[P] における分類子数が上限を超えた際に分類子の削除を行う．

[A] 内から適合度 F に基づくルーレット選択により二つの親を選択する．つまり，適合度が高い分類子ほど親として選択されやすい．次に，選択された二つの親を複製し，条件部の各ビットに対して，パラメータ χ と μ の確率で交叉と突然変異を実行する．XCS における交叉は，複製された二つの子の条件部の対応する順番の符号を交換する操作である．突然変異は，“0”，“1” であれば “#” に変換し，“#” であればその時の状態（State）の，対応する順番の値に変換する操作である．

これらの操作により生成された新たな子となる分類子は，[P] に追加される．このとき，[P] 内の分類子数が上限を超えた場合，分類子の削除を行う．削除される分類子は式 (7) によって算出される削除率に基づくルーレット選択により選ばれ，削除される．

$$deletionvote_j = as_j \times num_j \tag{7}$$

ここで，$deletionvote_j$ は削除率，as_j は重複度，num_j は重合度である．

4.5 ルールの最適化メカニズム

4.5.1 カバリング

カバリングとは，状態に照合される条件部を持つ新たな分類子を生成し，[P] に追加する操作である．[P] 内に状態（State）と照合する条件部を持つ分類子が存在しない，または [M] 内の行動選択肢がパラメータ θ_{nma} より少ない場合に実施される．また生成される分類子の条件部の符号は一定の確率 $P_\#$ で “#” に置き換えられる．これにより，学習初期のような状態空間の探索が不十分な場面においても環境適応が促進される．

4.5.2 包　　摂

包摂 (Subsumption) とは，行動集合 [A] において，他の分類子と同じ行動部を持ち，より一般化された条件部を持つ分類子が存在する場合に，これらを統合する操作である．具体的には，これらの分類子を集約し，重合度を足し合わせる．例えば，同じ行動集合 [A] 内に条件部が “##01” で重合度が 3 である分類子 A と，条件部が “#101” で重合度 2 である分類子 B が存在する場合，より一般化された分類子 A に分類子 B が集約される．その結果，条件部が “##01”，重合度が 5 である単一の分類子が形成される．包摂によってより一般的なルールが形成されることで，ルール集合における冗長性の低減が可能となる．

5 二つのルール集合を獲得する拡張 XCS

本節では，マルチエージェント迷路環境に向けて，状態表現およびルール集合を拡張した XCS について提案する．本研究では，他エージェントの存在を状態として明示的に扱うため，周囲 8 近傍を 2 ビットで表現する状態符号化において，従来は未使用であった “10” を他エージェントの表現として用いる．本研究で提案する拡張 XCS は，マルチエージェント迷路環境における行動決定を，競争的側面と協調的側面に分離して学習を行う．具体的には，自己の即時的なゴール到達を優先する競争的ルール集合と，チーム全体のゴール到達を優先する協調的ルール集合を持ち，各集合に対してルールを生成する．そして，学習時および評価時に，協調的ルール集合の協調的戦略に基づく行動選択を行う．これにより，競争的戦略および協調的戦略を表現しつつ，マルチエージェント迷路環境への適応を目指す．

5.1 行 動 選 択

標準的な XCS では，ε-greedy 法 [12] やルーレット選択 [11] による行動選択が採用されているが，本手法では，エージェントの行動選択に予測配列（[PA]）の値に基づくボルツマン選択 [12] を採用する．

ボルツマン選択は，各行動の予測報酬に応じた重みづけ確率を用いて行動を選択する．ボルツマン選択では，温度パラメータ T の調節によって，各行動が選択される確率を制御できるため，学習における行動のランダム性，すなわち学習時の探索率を柔軟に調節することが可能である．具体的には，行動 a_i を選択する確率を $Pr(a_i)$ とすると，その確率は式 (8) で計算される．

$$Pr(a_i) = \frac{e^{\frac{P(a_i)}{T}}}{\sum_{a_j} e^{\frac{P(a_j)}{T}}} \tag{8}$$

なお，e はネイピア数，T $(T > 0)$ は探索の度合いを制御する

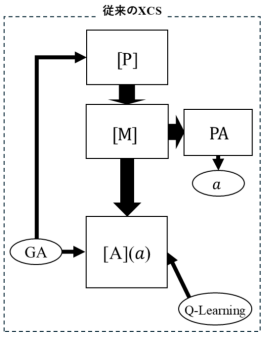

図 3　従来の XCS の処理の流れ

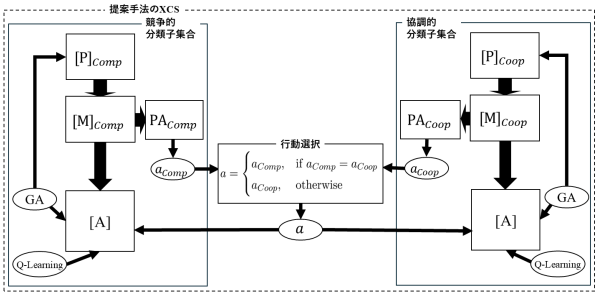

図 4　提案手法 XCS の処理の流れ

ための温度パラメータである．T が高い場合は，各行動の選択確率が均等化されて探索が促進され，$T \to 0$ に近づくにつれて最大予測報酬を持つ最適行動の選択確率が高まる．これにより，予測報酬に応じた行動の選択とパラメータを用いた容易な探索率の制御が可能である．

5.2　提案する拡張 XCS の構造と処理の流れ

標準的な XCS は，単一の分類子集合によって学習および行動決定を行う．一方で，提案する手法では，XCS が二つの分類子集合を持つように拡張する．標準的な XCS と提案する手法における学習のプロセスをそれぞれ図 3 と図 4 に示す．図中の $Comp, Coop$ はそれぞれ競争的（Competitive）および協調的（Cooperative）を意味し，[P]，[M]，PA，a, [A](a) はそれぞれ，分類子集合，照合集合，予測配列，提案する行動，および行動部に a を持つ分類子の行動集合を表す．なお，これらの図では，提案手法と直接的な関係がないカバリングや包摂に関する処理については省略している．提案手法の具体的な説明を以下に示す．

まず，提案手法における XCS は，標準的な XCS と同様に環境から状態を受け取る．受け取った状態に基づき，二つの分類子集合 $[\mathrm{P}]_{Comp}$ および $[\mathrm{P}]_{Coop}$ において，それぞれ照合集合 $[\mathrm{M}]_{Comp}$，$[\mathrm{M}]_{Coop}$ を生成する．次に，各照合集合から予測配列を作成する．行動選択には前述したボルツマン選択を採用する．その結果，各集合はそれぞれ独立した行動である a_{Comp}, a_{Coop} を提案する．

エージェントが実際に環境に対して実行する行動 a は，二つの分類子集合から提案された行動 a_{Comp}, a_{Coop} の比較に基づいて決定される．両集合の提案が一致する，すなわち $a_{Comp} = a_{Coop}$ の場合には，その行動を採用する ($a = a_{Comp} = a_{Coop}$)．一方で，両集合が提案する行動が不一致，すなわち $a_{Comp} \neq a_{Coop}$ の場合には，協調的ルール集合が提案した行動を採用する ($a = a_{Coop}$)．この行動決定により，協調的分類子集合に基づく行動が優先的に選択され，複数エージェントによるチームとしてのゴール到達に貢献できる行動を学習させる．

強化学習によるパラメータ更新や遺伝的アルゴリズムによる分類子の生成と削除は，標準的な XCS と同様であり，4.3 節および 4.4 節で述べた通りである．

5.2.1　競争的分類子集合

二種類の分類子集合のうち競争的分類子集合について述べる．競争的ルール集合は，各エージェントが他エージェントの存在を考慮せず，自身の即時的なゴール到達を目的とした行動の学習を目標としている．つまり，短期的な報酬獲得を重視した貪欲な行動方策の獲得を目指して設計する．

具体的には，ボルツマン選択の温度パラメータ T を低く設定する．温度パラメータ T を低くすることで探索率が抑制され，学習過程で既に高い予測値と適合度を持つ分類子に基づく行動が優先的に選択される．その結果，エージェント自身にとって即時的に有利な経路を選択することが期待される．また，競争的分類子集合では割引率 γ を低めに設定する．割引率を低くすることで，将来得られる報酬よりも現在の報酬を重視する学習が促進され，長期的な報酬の最大化よりも，短期的なゴール到達を優先する行動の学習が期待できる．

以上のように，競争的分類子集合では，探索率 T および割引率 γ をともに低く設定することで，エージェントが環境との相互作用を通じて，短期的かつ自己中心的な行動方策を安定して学習することを目的とする．

5.2.2　協調的分類子集合

協調的分類子集合では，各エージェントが即時的な報酬の最大化ではなく，他エージェントの行動や将来的な状態遷移を考慮した行動選択を目指す．

競争的分類子集合とは異なり，温度 T と割引率 γ を高めに設定する．高めの温度パラメータ T の設定により，行動選択の確率分布がより均一となり，多様な行動が選択されやすい状態となる．これにより，単一の最短経路に偏らず，チーム全体として有効な行動の探索が期待できる．また割引率 γ においても高めに設定し，将来的に得られる報酬を重視した学習を行う．これにより，短期的には不利に見える行動であっても，将来的に全エージェントの目標達成につながる行動が，競争的分類子集合と比較して評価されやすくなる．

協調的分類子集合の学習においては，シングルエージェントのゴール到達だけでなく，チーム全体のゴール到達を最優先とした報酬設計を行う．本手法の協調的分類子集合では，自身がゴール到達した際の報酬値をもう一方のエージェントの状態に依存して決定する条件付きの報酬として設計する．具体的には，両エージェントがゴールに到達した場合の報酬を R_{both}，自身のみがゴールに到達した場合の報酬を R_{self} と定義し，$R_{both} > R_{self}$ を満たすように報酬設計を行った．

以上の設定により，協調的分類子集合では，個々の即時的な

最適化よりも，マルチエージェント環境における長期的かつ全体的な目標達成を重視した行動戦略の獲得が期待できる．

6 迷路環境における評価実験

6.1 実験の内容

実験では，マルチエージェントに拡張した迷路問題において，提案手法の有効性を検証する．比較対象には，状態表現をマルチエージェント迷路探索用に変更した，ε-greedy 選択およびルーレット選択を用いた XCS を採用する．

ε-greedy 法 [12] は，確率 $\varepsilon(0 < \varepsilon < 1)$ でランダムな行動を選択し，確率 $1-\varepsilon$ で予測報酬が最も高い行動を選択する．行動 a_i を選択する確率を $Pr(a_i)$ とすると，その確率は式 (9) で計算される．

$$Pr(a_i) = \begin{cases} 1-\varepsilon, & \text{if } a_i = \arg\max_{a_j} P(a_j) \\ \frac{\varepsilon}{|A|-1}, & \text{otherwise} \end{cases} \tag{9}$$

ここで，$\varepsilon, P(a_i), |A|$ はそれぞれ，確率 ε，行動 a_i を選んだ時の予測報酬，エージェントの行動選択肢数である．

ルーレット選択 [11] は，分類子が提案する行動の予測報酬に比例した確率で行動を選択する手法である．ルーレット選択による行動選択確率は，各行動 a_i に対して，予測報酬を $P(a_i)$ とすると，式 (10) で表される．

$$Pr(a_i) = \frac{P(a_i)}{\sum_{a_j} P(a_j)} \tag{10}$$

実験では，1 回の学習（iteration）において，学習フェーズにおける環境の探索と，評価フェーズにおける獲得したルール集合の検証を行う．学習フェーズでは，強化学習によるパラメータの更新と遺伝的アルゴリズムによる分類子集合の更新を行うが，評価フェーズでは，分類子のパラメータの更新を行わず，予測報酬が最も高い行動が実行される．

評価指標として，ゴール到達までの平均ステップ数を用いる．これは学習性能を示す指標であり，エージェントが最適経路へ収束しているかどうかを評価するために用いる．また，XCS が学習したルールを比較するために，分類子集合から算出した行動ごとの予測報酬を用いる．分類子集合から算出した行動ごとの予測報酬は，図 5 のセルごとの値を可視化した図を用い，分類子集合がどのような行動を提案しているかを確認する．

6.2 実験環境およびパラメータ

本実験では，二体のエージェントが同一の迷路環境内でそれぞれゴール到達を目指すマルチエージェント経路探索を実施する．使用する迷路環境は，図 1 の Maze6 を参考に，新たに図 5 の迷路を作成した．図 5 の白色のセル，灰色のセル，黒色のセルはそれぞれ通路，障害物，ゴールを表している．迷路内には，座標 (1, 7) および (4, 4) に配置された二か所のゴールが存在し，前者をゴール 1(G1)，後者をゴール 2(G2) とする．エージェントのスタート位置は (1, 2)，(7, 2) とし，それぞれエージェント 1，エージェント 2 とする．この環境においては，両

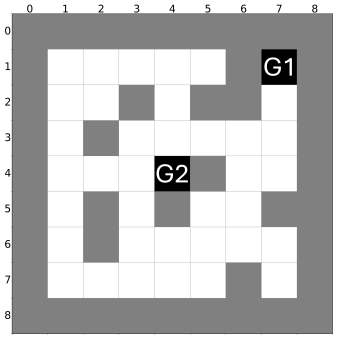

図 5 マルチエージェント迷路探索用 Maze6

表 2 両手法共通のパラメータの値と説明

パラメータ	値	説明
N	4,000	分類子数の最大サイズ
α	0.1	絶対的な正確さ κ の調節用パラメータ
β	0.2	学習率
ϵ_0	0.1	予測誤差の許容閾値
ν	5	絶対的な正確さ κ の調節用パラメータ
χ	0.6	交叉の発生確率
μ	0.02	突然変異の発生確率
θ_{GA}	100	GA の実行閾値

エージェント共にゴール 2 へ向かう経路が最短となり，シングルエージェント環境であればそれが最適行動に相当する．しかし，他方のエージェントが存在するセルには移動できない制約を設けており，片方のエージェントがゴールに到達している場合，もう一方のエージェントはそのゴールへ遷移することができない．その結果，一方のエージェントは，より遠方に位置するゴール 1 へ向かう必要が生じる．

いずれかのエージェントがゴールに到達した場合，当該エージェントはその時点で学習および移動を停止し，残る一方のエージェントが探索を継続する．両エージェントがいずれかのゴールに到達する，もしくはエージェントの移動ステップ数が 50 ステップに達した時点で 1 回の学習を終了する．この学習を 100,000 回実施する．従来の Maze6 における実験と同様に，各エージェントは周囲 8 近傍の状態を観測可能であり，観測方向と対応した 8 種類の移動を選択できる．報酬は，各エージェントがゴールに到達した場合にのみ，当該エージェントに与えられる．

実験における XCS のパラメータ設定は，文献 [14] を参考に，表 2 のように設定した．加えて，エージェントに与える報酬を表 3 のように設定した．また，割引率および探索率は，それぞれ表 4，表 5 のように設定した．なお，γ，χ，μ は本研究のマルチエージェント迷路環境に向けて設定したため，文献 [14] とは異なる．

6.3 実験結果

図 6 は，学習回数に対する両方のエージェントがゴールに到達するまでのステップ数を表している．横軸は iteration を表し，iteration は学習と評価からなる 1 サイクルを意味する．縦軸はゴール到達までのステップ数（Steps to Food）を表し

表 3　各分類子における報酬設定

	従来 XCS	競争的分類子集合	協調的分類子集合
1 体のエージェントのみゴール	1,000	1,000	500
両エージェントがゴール	1,000	1,000	1,000

表 4　割　引　率

パラメータ	値	説明
γ_{Comp}	0.65	競争的分類子集合の割引率
γ_{Coop}	0.85	協調的分類子集合の割引率
γ	0.70	従来 XCS の割引率

表 5　探　索　率

パラメータ	値	説明
T_{Comp}	50	競争的分類子集合の温度
T_{Coop}	100	協調的分類子集合の温度
ε	0.3	従来 XCS(ε-greedy) の探索率

ており，値が小さいほど短いステップ数でゴールに到達できることを意味する．本図では，100 回ごとの平均値を一つの値とし，同一条件の実験を 3 回実行した結果について，それらの平均値をプロットしている．青の線はルーレット選択，緑の線はε-greedy 選択，黄の線は提案手法の結果を示す．また，赤の点線は最適行動におけるゴール到達ステップ数である 6 を示している．

図 7，図 8，図 9，図 10 には，それぞれ，ε-greedy 選択を採用した XCS，ルーレット選択を採用した XCS，競争的ルール集合，協調的ルール集合が保有する分類子に基づき，各セルにおける行動を可視化した結果を示す．

図 7，図 8，図 9，図 10 では，図 5 の空白を白色，T を灰色，F を黒色に色付けし，各セルの赤い矢印はそのセルにおける最も予測報酬が高い行動，青い矢印はその他の行動を示し，青の矢印では濃淡が濃いほど予測報酬が高い行動を示す．赤色の正方形で囲われたセルはそのエージェントの学習時および評価時のスタート地点を示す．また，図 9 における，黄色，緑色，水色の正方形で囲われたセルは，6.4 節で述べる分析において説明するセルを示したものである．図中のゴールのセル (黒セル) 内に記されている白文字の G1, G2 はそれぞれゴール 1，ゴール 2 を意味している．これらの図を用いて，それぞれの手法が獲得した行動の特徴，特に競争的ルール集合が示す行動と協調的ルール集合が示す行動を比較する．なお，左右の図のうち左がエージェント 1 が学習した分類子に基づく行動，右がエージェント 2 が学習した分類子に基づく行動を示している．

6.3.1　平均ステップ数

図 6 より，ルーレット選択を採用した XCS は，ゴール (Food) までのチームの平均ステップ数が約 10 ステップから 35 ステップの範囲で大きく変動していることがわかる．一方で，ε-greedy 選択を採用した XCS におけるゴールまでのチームの平均ステップ数は，約 10 ステップから 20 ステップの範囲に収まっている．また，提案手法は ε-greedy 選択と同程度の推移を示しているものの，40,000 回の学習以内において，ε-greedy 選択を採用した XCS と比較してゴールまでの平均ステップ数がわずかに短くなる傾向がみられる．

6.3.2　学習した行動

図 7 から，ε-greedy 選択によって学習される行動は，特定のゴールが過度に選択される傾向がみられる．具体的には，エージェント 1 はゴール 2 へ向かい，エージェント 2 はゴール 1 へ向かう行動を提案している．特にエージェント 2 の場合，ゴール 2 周辺のセルであってもゴール 2 へ向かわず，ゴール 1 に向かう行動が強調されている．図 8 から，ルーレット選択により学習される行動は，ε-greedy 選択と比較して両方のゴールに向かう行動が多く提案されていることが確認できる．しかし，エージェント 1 およびエージェント 2 の双方がゴール 2 へ向かう行動を選択しており，協調が十分に行われていない点が問題として挙げられる．

提案手法である図 9 および図 10 を確認すると，エージェント 2 がゴール 2 へ向かい，エージェント 1 がゴール 1 へ向かう，つまりエージェント 1 はゴール 2 をエージェント 2 に譲る結果となった．この行動は，チームとして最小のステップ数でゴールへ到達できるため，チームとして最適な行動である．しかし，ゴール到達に必要な行動に関して，競争的ルール集合と協調的ルール集合の内容に，類似した傾向がみられた．エージェント 1 が学習した，競争的ルール集合および協調的ルール集合において，それぞれ最も予測報酬が高い行動を確認すると，いずれの集合においてもゴール 1 に向かう行動が選択されている．また，競争的ルール集合は，ゴール 2 の周辺のセルにおいても，ゴール 1 に向かう行動を選択している確認された．

次に，エージェント 2 が学習した競争的ルール集合と協調的ルール集合が示す最も予測値が高い行動を比較する．競争的ルール集合においては，(7, 2)，(6, 3)，(5, 3) といったセルを経由してゴール 2 へ到達する行動が示されており，ゴール 2 へ最短で到達する行動が選択されていることが確認できる．一方で，協調的ルール集合においては，ゴール 2 に隣接する (3, 5) セルにおいて，ゴール 2 に向かう行動ではなく，ゴール 1 に向かう行動が選択されていることが確認できる．また，(7, 2)，(6, 3)，(6, 4) セルといったように，ゴール 2 に最短で移動するための行動ではなく，ゴール 1 に移動するための行動を考慮した行動が選択されており，結果としてゴール 2 の右下方向のセルへ移動する行動が選択されていることが確認できる．

6.4　考　　察

提案手法の競争的ルール集合および協調的ルール集合が獲得したルールの分析と，平均ステップ数の分析を行う．

6.4.1　提案手法のエージェント 1 が学習した行動の分析

6.3.2 節で示したように，エージェント 1 において，本来は競争的ルール集合として学習されるべき行動であるにもかかわらず，遠いゴール (ゴール 1) へ移動するといった，提案手法におけるルール獲得の意図とは異なる学習が行われていることがわかった．そこで，競争的ルール集合のルールを分析する．具体的には，エージェント 1 のスタート位置に近く，かつゴール

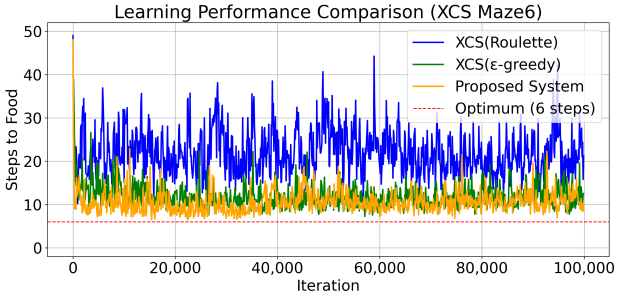

図 6　学習回数ごとのゴールまでの平均ステップ数

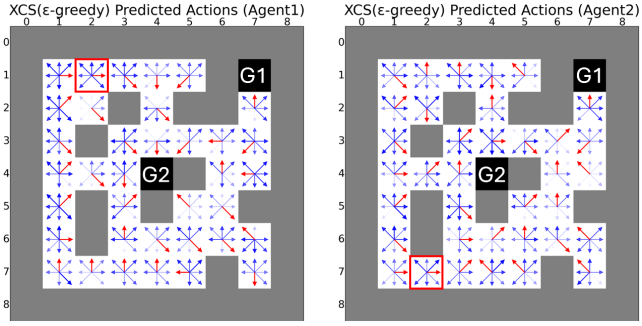

図 7　ε-greedy 選択を採用した XCS が学習した行動

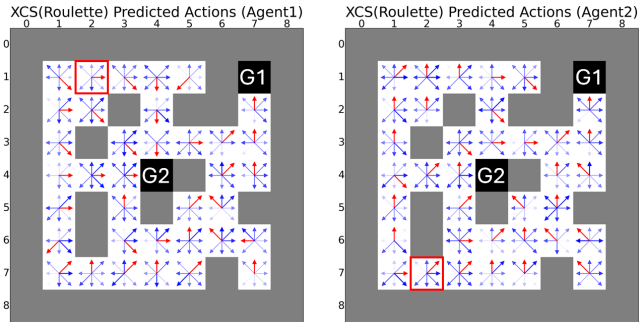

図 8　ルーレット選択を採用した XCS が学習した行動

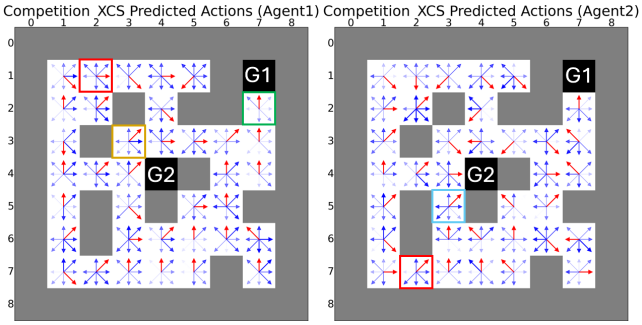

図 9　競争的ルール集合が学習した行動

2 の隣接セルである，図 9 の左図内のセル (3, 3)，(3, 4)，(3, 5) のセルごとの予測報酬の値，分類子について分析する．各移動方向に対する予測報酬の値を表 6 に示す．この表は移動方向に対しての予測報酬の値をセルごとに示したものである．太字の数値はそのセルにおける最も予測報酬が高い値を持つ移動方向である．この表から，どのセルにおいても，ゴール 1 の方向である右方向の予測報酬が高く，ゴール 2 の方向の予測報酬が

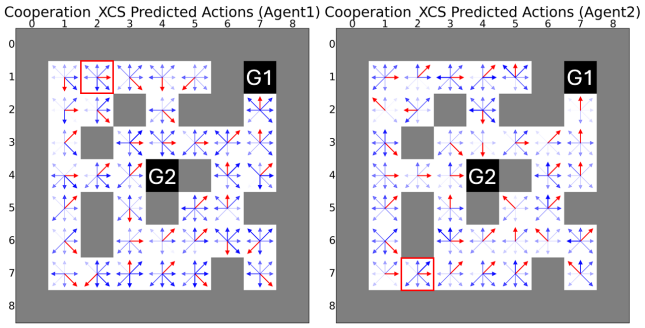

図 10　協調的ルール集合が学習した行動

表 6　エージェント 1 の競争的ルール集合による方向ごとの予測報酬

方向	G2 左上セル (3, 3)	G2 上セル (3, 4)	G2 右上セル (3, 5)
↑	110.75	208.36	273.91
↗	**177.44**	159.41	260.33
→	172.78(G1 方向)	**273.74**(G1 方向)	**422.16**(G1 方向)
↘	112.93(G2 方向)	175.04	258.56
↓	95.57	152.64(G2 方向)	272.57
↙	83.87	87.05	249.75(G2 方向)
←	100.82	109.43	133.87
↖	77.34	171.82	165.51

表 7　エージェント 1 の競争的ルール集合のセル (3, 3) に照合する右下 (ゴール 2 方向) の行動部を持つ分類子

condition	action	fitness	prediction
#10####1##00#1#0	3	0.249028	71.3532362
#1#0001#######0#	3	0.191684	110.4063601
01000###00#0##00	3	0.167230	178.1726725
###0001###0#0###	3	0.021383	111.0134999
####0#1#0##0##0#	3	0.012649	110.6352085

表 8　エージェント 1 の競争的ルール集合のセル (2, 7) において上 (ゴール 1 方向) の行動部を持つ分類子

condition	action	fitness	prediction
11##############	0	0.999870	1000

低いことが確認できる．

次に，エージェント 1 の競争的ルール集合が獲得した分類子の代表例として，セル (3, 3) において右下方向の行動を示す分類子を分析する．比較対象として，エージェント 1 の競争的ルール集合における，ゴール 1 に隣接するセルであるゴール 1 直下のセル (2, 7) において上方向の行動を示す分類子と，エージェント 2 の競争的ルール集合における，ゴール 2 左下のセル (5, 3) において右上方向の行動を示す分類子を用いる．また，セル (3, 3)，(2, 7)，(5, 3) は，図 9 で，それぞれ黄色，緑色，水色の正方形で囲われたセルである．これらの分類子のうち，適合度 (fitness) が上位の 5 件を，それぞれ表 7，表 8，表 9 に示す．ただし，表 8 に該当する分類子は一つしか存在しなかったため，その分類子のみ記載している．

表 7 を確認すると，全体的に prediction(予測値) が，表 8，表 9 の分類子と比較して非常に低いことがわかる．また，表 8 および表 9 に示した分類子では，ゴール方向に対応する condition（条件部）のビット列が “11” と明示的に記述されているものが

表 9　エージェント 2 の競争的ルール集合のセル (5, 3) において右上 (ゴール 2 方向) の行動部を持つ分類子

condition	action	fitness	prediction
0#110##0##0#0100	1	0.436209	1000
00110##0##0#0100	1	0.157684	1000
00#1##000###0100	1	0.149729	1000
00#10##0####0100	1	0.069038	1000
0#11##000#0#0100	1	0.069038	1000

多く確認できる．一方，表 7 に示した分類子では，“1#” のように第 1 ビットのみが固定され，第 2 ビットがワイルドカード “#” で一般化されているものが多く含まれている．ここで，condition の各 2 ビットは観測方向の状態を表しており，“11” はゴール，“10” は他エージェントを意味する．例えば，表 8 に示した分類子「11##############」は，直上方向にゴールが存在する状態を明示的に表現しており，その他の方向は一般化されている．このように，ゴールの存在を特定した条件部を持つ分類子は，特定のゴール隣接状態に強く対応している．これに対して，“1#” という表現は，第 1 ビットが 1 であることのみを指定しており，観測方向の対象がゴール（11）である場合と他エージェント（10）である場合の双方を包含する．したがって，ゴールと他エージェントを区別しない条件部となっている．

このような一般化の違いにより，エージェント 1 の競争的ルール集合では，ゴール 2 隣接状態に特化した分類子が十分に形成・保持されていないことが示唆される．その結果，該当セルにおける予測値が相対的に低くなり，安定的な強化が行われにくい状況が生じていると考えられる．そして，予測値の低さから，エージェント 1 においては，ゴール 2 への到達に対応するルールが安定的に保持されず，生成と削除が繰り返されていることが示唆される．この要因としては，他エージェントの行動の影響により当該ルールの予測誤差が一時的に増大し，適切に強化されにくい状況が生じていることが考えられる．一方で，エージェント 2 の競争的ルール集合が保有する表 9 の分類子は，非常に高い予測値を持っていることから，エージェント 2 は学習時にゴール 2 に高頻度で到達したことが示唆される．さらに，ゴール方向のビットが一般化されていることから，この分類子集合はゴール 2 をゴール “11” と認識するのではなく，他エージェント “10” もしくはゴール “11” と認識している．エージェントは “11” である可能性を持つ方向を選択しても，実際には “10” だった場合，報酬を受け取ることができないため，予測値が小さくなったことがわかる．以上から，エージェント間の学習機会を均等化する方策の検討が必要である．また，学習初期などゴール到達経験が十分でない段階では一般化を抑制し，ゴール到達に寄与するルールが一定程度形成された後に一般化を促進する段階的な制御方法の導入も重要である．特にマルチエージェント環境では，他エージェントの行動の影響により分類子の予測誤差が一時的に増大することがあり，その結果として不適切な包摂が生じ得る．したがって，一般化および包摂の進行を学習状況に応じて適切に制御する機構の導入が求められる．

6.4.2　提案手法のエージェント 2 が学習した行動の分析

図 9 および図 10 の右図の結果，つまりエージェント 2 が学習した行動について分析する．競争的ルール集合においては，図 9 に示すように，ゴール 2 へ到達するまでのステップ数は 3 ステップであり，エージェント 2 にとって最も短時間で報酬を獲得できる行動が学習されていることが確認できた．これは，競争的ルール集合において，探索率および割引率を協調的ルール集合と比較して低い値に設定したことが影響している．割引率の値が低い場合，短期的に得られる報酬を重視した学習が行われるため，可能な限り早くゴールに到達する行動が選択されやすくなる．また探索率についても，低い値を設定しているため，一度報酬を得ることができた行動をより重視し，それ以外の行動を選択しにくくなる．この結果，競争的ルール集合は，ゴール 2 へ向かう最短経路に対応した行動を優先的に学習したと考えられる．一方で，図 10 の右図に示すエージェント 2 の協調的ルール集合においては，ゴール 2 へ向かう行動を学習しているものの，競争的ルール集合と比較して，ゴール 1 への移動に向けた行動も学習されていることが確認できた．これは，競争的ルール集合とは対照的に，探索率および割引率を高い値に設定したことが影響している．具体的には，割引率を高く設定したことで，短期的な報酬のみならず，将来的に得られる報酬も考慮した学習が行われる．さらに，探索率が高いことから，ゴール 2 へ向かう行動に固執することなく，それ以外の行動も選択されやすくなる．その結果として，協調的ルール集合では，ゴール 1 を含む複数のゴールを考慮した行動が獲得されたと考えられる．既に述べたように，エージェント 2 はエージェント 1 と比較してゴール 2 に関する学習機会が多く，競争的ルール集合と協調的ルール集合で異なる方針のルールが生成されたことが示唆される．

6.4.3　平均ステップ数に関する考察

提案手法を用いた結果，他の XCS と比較してゴール到達までの平均ステップ数が短縮されたことがわかった．要因として以下の二点が挙げられる．一つ目の要因として，協調的ルール集合による長期的な報酬を重視した学習が挙げられる．この学習方針により，エージェントは直近で到達可能なゴールに固執することなく，状況に応じて適切なゴールを選択する行動を獲得した．二つ目の要因として，評価時における行動選択の方法の影響が挙げられる．提案手法では，評価時に競争的ルール集合と協調的ルール集合が異なる行動を提案した場合に，協調的ルール集合によって獲得されたルールを優先して選択する．この仕組みにより，一方のエージェントが遠いゴールへ向かう行動を選択することで，チーム全体のゴール到達までのステップ数が短縮された．以上の傾向は上野らによっても理論的に示されており，提案手法の有効性を裏付けている [15].

7　おわりに

本研究では，解釈可能なルール集合を獲得する XCS を用いたマルチエージェント迷路環境の学習を通したルールの獲得を

目的とし，自らの即時的なゴール到達を優先する競争的ルール集合と，チーム全体のゴール到達を優先する協調的ルール集合という二つのルール集合に基づいて学習を行う拡張 XCS を提案した．提案手法では，探索率，割引率，および報酬設計を調節することにより，競争的ルール集合および協調的ルール集合それぞれに適したルールの生成を行う．そして，学習時および評価時において，これらの二つのルール集合のうち，チーム全体のゴール到達を優先する協調的ルール集合に基づいて行動を選択することで，XCS によるマルチエージェント迷路環境の学習の実現を目指した．

評価実験では，2 体のエージェントと 2 つのゴールが存在する Grid Worlds 型の迷路環境において，ε-greedy 選択およびルーレット選択を採用した標準的な XCS と提案手法の学習結果を比較した．その結果，提案手法は，標準的な XCS と比較して，ゴール到達までのチームの平均ステップ数がより早く収束する傾向を示した．一方で，エージェント間においてゴールごとの到達経験に偏りが生じたことが原因で，本来は自らの即時的なゴール到達を優先する競争的ルール集合においても，遠いゴールへ向かう行動を選択するルールが学習される場合が確認された．そのため，競争的ルール集合と協調的ルール集合において，類似した行動を選択するルールが生成されるという課題が明らかとなった．

今後の課題としては，競争的ルール集合と協調的ルール集合の間で，より明確な差異を有するようにルールを生成することが挙げられる．特に，エージェント間でゴールごとの到達経験に偏りが生じることが，競争的ルール集合においても協調的な行動を選択するルールが学習される要因となったため，エージェントごとの学習機会を均等にする方法について検討したい．

謝　　辞

本研究の一部は，科学研究費基盤研究 (B)（課題番号 JP24K03001，JP25K03227）の援助による．

文　　献

[1] S. R. Islam, W. Eberle, S. K. Ghafoor and M. Ahmed, “Explainable Artificial Intelligence Approaches: A Survey,” *arXiv preprint arXiv:2101.09429*, 2021.

[2] J. H. Holland, Escaping Brittleness: The Possibilities of General Purpose Learning Algorithms Applied to Parallel Rule-based System, “Machine Learning,” *Evolutionary Computation*, vol. 2 pp. 593–623, 1986.

[3] S. W. Wilson, “Classifier Fitness Based on Accuracy,” *Evolutionary Computation*, vol. 3, no. 2, pp. 149–175, 1995.

[4] P. L. Lanzi, “An Analysis of the Generalization in the XCS Classifier System,” *Evolutionary Computation*, vol. 7, no. 2, pp. 125–149, 1999.

[5] P. L. Lanzi and S. W. Wilson, “Toward Optimal Classifier System Performance in Non-Markov Environments,” *Evolutionary Computation*, vol. 8, no. 4, pp. 393–418, 2000.

[6] P. L. Lanzi, “An Analysis of Memory Mechanism of XCSM,” *Genetic Programming*, vol. 98, pp. 643–651, 1998.

[7] M. V. Butz, *Anticipatory Learning Classifier Systems*, vol. 4, Springer US, Boston, MA, USA, 2002.

[8] A. Siddique, W. N. Browne and G. M. Grimshaw, “Frames-of-Reference-Based Learning: Overcoming Perceptual Aliasing in Multistep Decision-Making Tasks,” *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 1, pp. 174–187, 2022.

[9] F. Uwano and W. N. Browne, “Enhancing XCS with Dual-Stream Identification for Perceptual Aliasing in Multi-Step Decision-Making,” *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2025)*, pp. 499–507, 2025.

[10] 北野 宏明, “遺伝的アルゴリズム,” 産業図書, vol. 1, pp. 26–35, 1992.

[11] M. V. Butz and S. W. Wilson, “An Algorithmic Description of XCS,” *Journal of Soft Computing*, vol. 6, pp. 144–153, 2002.

[12] R. S. Sutton and A. G. Barto, “Reinforcement Learning: An Introduction Second Edition,” MIT Press, 2018.

[13] C. J. C. H. Watkins and P. Dayan, “Technical Note Q-Learning,” *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[14] M. V. Butz, T. Kovacs, P. L. Lanzi and S. W. Wilson, “Toward a Theory of Generalization and Learning in XCS,” *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 1, pp. 28–46, 2004.

[15] F. Uwano, N. Tatebe, Y. Tajima, M. Nakata, T. Kovacs and K. Takadama “Multi-Agent Cooperation Based on Reinforcement Learning with Internal Reward in Maze Problem,” *SICE Journal of Control, Measurement, and System Integration*, vol. 11, no. 4, pp. 321–330, 2018.