2C-06

チャネルマスキングによるブラックボックスモデルの解釈性向上

飯島 敏也† 大西 直† 土井 護† 杉原 堅也†
三菱電機株式会社†

1.はじめに

対戦ゲームの環境において，機械学習を用いて相手の行動を分析し，自動で相手の作戦を判断することは有用である．これにより，相手の行動に対抗する適切な作戦を立てることが可能となる．しかし，機械学習モデルがなぜその判断をしたのか人間が解釈することは困難である．

機械学習モデルに入力した特徴量の寄与度を算出することで，判断根拠を可視化することにより，解釈を補助する手法がある [1]．ランダムな画素にマスクした入力画像をモデルに入れ，予測確率の変化量を求める．これを繰り返すことで，入力した画像の各画素の寄与度を算出できる．しかし，画像の画素やチャネルには物体に関する情報がないため，寄与度の分布をもとに人間が解釈を行う必要がある．また，ランダムにマスクするため，判断根拠の可視化には多数の試行が必要である．

本稿では，対戦ゲームなどの作戦を予測する機械学習モデルの判断根拠の解釈性向上を目的とする．対戦環境情報を，位置や行動などの情報を持つチャネルに分け，特定のチャネルをマスクして寄与度を算出するチャネルマスキングにより，解釈性取得までの試行回数を削減しながら解釈性を向上する手法を提案する．

2.検証環境

機械学習による相手の行動分析の簡易的な検証環境としてミツバチ対スズメバチの対戦ゲームを作成した(図 1)．このゲームは Griddly [2]上で動作する．ミツバチは花から資源を集め，スズメバチはミツバチまたは巣を襲撃して資源を奪い，資源の総量を競う．各エージェントは，ルールベースに基づいて行動する．スズメバチには 3 つの作戦に基づいた行動パターンで対戦を行う．対戦経過から，スズメバチがどの作戦を取っているか分類する．

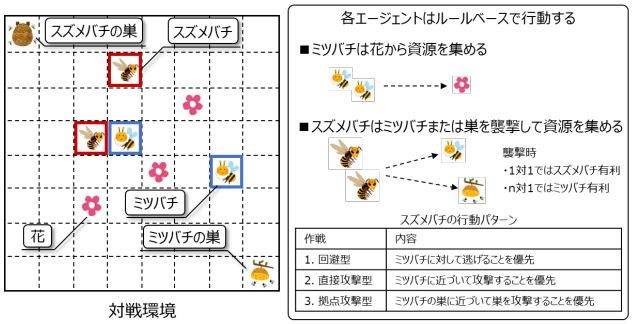

図 1 ミツバチ対スズメバチ

3.提案手法

図 2に本手法を用いた解釈性取得の流れを示す．

3.1. チャネルマスキング

入力された対戦環境を特徴量マップに変換する．特徴量マップは，位置や行動に関する情報を 2 値画像化する．位置については，エージェントの種類ごとに，存在する場合は 1，存在しない場合は 0 とする．また，行動については，対象の行動をした場合は 1，していない場合は 0 とする．これにより，チャネルごとに対戦環境の情報を持った特徴量マップを得られる．

次に，特定の特徴量マップに対してマスクを付与する．ここで，特徴量マップは，チャネルごとに行動や位置などの情報を持つため，特定の情報にマスクできる．マスクの方法としては，単一もしくは複数のチャネルの情報を全て 0 にする．

3.2. 解釈性の取得

3.1節により生成した特徴量マップを入力として教師あり学習により学習済みモデルを作成する．学習済みモデルに対して，特定のチャネルにマスクを付与した特徴量マップを入力し，マスクの有無による各クラスの正解率を比較し寄与度を算出する．チャネルの寄与度により，チャネルに関連した対戦環境の情報が分類結果にどのような影響を与えるかを解釈することができる．

Channel Masking for Explanation of Black-box Models
† Toshiya Iijima, Tadashi Onishi, Mamoru Doi, Kenya Sugihara, Mitsubishi Electric Corporation

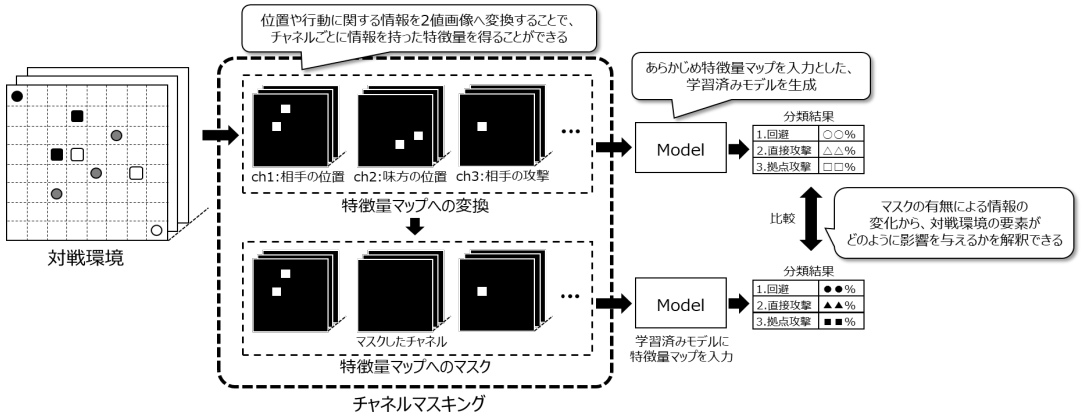

図 2 本手法の流れ

4.実験

本手法の有効性を検証するために，2章で述べた検証環境を用いて実験を行った．3DCNN [3]を用いてスズメバチの行動パターンを分類するモデルを作成し本手法を適用した．各行動パターンに影響しうる情報に着目し，次の 3 パターンでマスクの付与を行い(a)マスクなしの場合と比較した．

(b)ミツバチに関するチャネルのマスク
　(ミツバチの位置，ミツバチへの攻撃)
(c)ミツバチの巣に関するチャネルのマスク
　(ミツバチの巣の位置，ミツバチの巣への攻撃)
(d)スズメバチの移動に関するチャネルのマスク
　(スズメバチの移動，スズメバチの移動方向)

特徴量マップへのマスクの有無が，学習済みモデルの正解率に与える影響を評価した．

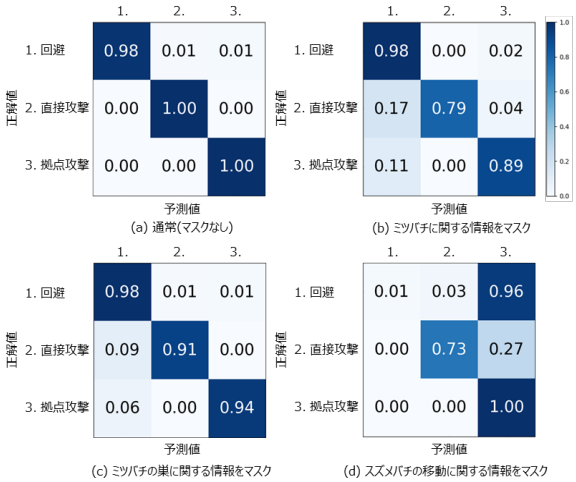

図 3 マスクの有無による分類結果の違い

5.結果

図 3に特徴量マップへのマスクの有無による分類結果の混同行列を示す．図 3 (b)の場合，直接攻撃型の正解率が低下した．図 3 (c)の場合，あまり変化が見られなかった． 図 3 (d)の場合，回避型の正解率が大幅に低下した．また，直接攻撃型の正解率も低下した．これにより各クラスに対するチャネルの寄与度がわかる．

これらの結果から，(b)の情報は直接攻撃の分類に寄与する．(c)の情報は，全体的な分類にあまり寄与しない．(d)の情報は回避型と直接攻撃型であるかの分類に寄与する．これにより各クラスの分類に対して重要となる対戦環境の情報がわかるため，解釈性を向上することができた．

6.おわりに

本稿では，機械学習モデルの判断根拠の解釈性向上のためにチャネルマスキング手法を提案した．対戦環境を特徴量マップに変換することで，チャネルごとに情報を持った特徴量を得ることができ，解釈性を得るまでの試行回数を減らすことができる．

検証環境を用いて，特定の特徴量マップへのマスクの有無による分類結果を比較した．チャネルの分類結果への寄与度により，各クラスの分類に対して重要となる対戦環境の情報がわかるため，解釈性を向上することができた．

今後は，より複雑な環境や実世界の問題において，提案手法の有効性を検証する予定である．

参考文献

[1] V.Petsiuk, A.Das , K.Saenko, Rise: randomized input sampling for explanation of black-box models, British Machine Vision Conference (BMVC), 2018.

[2] C. Bamford, Griddly: A platform for AI research in games, Software Impacts, 2021.

[3] S. Ji, W. Xu, M. Yang , K. Yu, 3D Convolutional Neural Networks for Human Action Recognition, IEEE transactions on pattern analysis and machine intelligence, 2012.