

二重過程理論による説明可能な AI 開発手法の提案

内田 輝[†] 松原 正樹^{††} 若林 啓^{††} 森嶋 厚行^{††}

[†] 筑波大学 情報メディア創成学類 〒 305-8550 茨城県つくば市春日 1-2

^{††} 筑波大学 図書館情報メディア系 〒 305-8550 茨城県つくば市春日 1-2

E-mail: [†]hikaru.uchida.2020b@mlab.info, ^{††}{masaki,kwakaba}@slis.tsukuba.ac.jp,

^{†††}morishima-office@ml.cc.tsukuba.ac.jp

あらまし 説明可能な AI システムの開発手法の研究は、深層学習分野において、今日の注目を集めている話題の一つである。本論文では、心理学における二重過程理論に着想を得て、論理的に推論を行い、理解可能な説明を出力するシステム 2-AI を開発するための手法を提案する。提案手法では、多クラス分類の各クラスについて、人間が理解可能な特徴をクラウドワーカーに入力してもらい、クラスと特徴の対応表を作成する。各特徴に関連する訓練データを収集し、これら集めたデータと対応表を基に、各特徴の有無を推定する深層学習モデルを複数生成する。各深層学習モデルの推定結果を組み合わせることで、分類対象が特定のクラスに属する理由を説明することが可能なシステム 2-AI を作成する。実験では、アジアの国旗の画像分類に本手法を適用し、その有効性と課題を検討した。

キーワード 説明可能 AI, クラウドソーシング, 二重過程理論

1 はじめに

今日、深層学習分野では、説明可能な AI の開発手法（例：[1][2][3]）が注目を集める話題の一つとなっている。AI の普及に伴い、人間社会、特に倫理、医療、法律のような AI の誤判定が人間に対し大きな影響を与える分野で、AI が誤判定をした場合の説明責任が求められるようになってきた。特に近年は、自動運転車の誤判定が原因となって発生する事故¹件数が増加しているため、AI に説明可能性を求める動きはより顕著になってきている。

人間は深層学習モデルがなぜその推定結果を出力したのか理解できないことがある。なぜなら深層学習では、学習過程がブラックボックスであり、その推定基準は人間が読み解くには複雑である場合があるためである。

本研究では、説明可能な AI を構築するために、心理学の二重過程理論 [4] に着想を得たヒューマン・イン・ザ・ループの手法を提案する。二重過程理論では、高速で直感的なシステム 1 と、低速で論理的なシステム 2 の 2 つのシステムで人間の思考が構成されると仮定している。我々は、システム 1 とシステム 2 のそれぞれに基づいた 2 つの AI を組み合わせることで、人間が理解可能な説明を効率的に生成することが可能であると推察する。本研究では、理想的な説明とは、論理的に書かれた、正確で、非冗長な説明であり、人間が比較的容易に理解可能な言葉で綴られた説明であると仮定する。

提案手法では、心理学における二重過程理論のシステム 1 とシステム 2 を、2 つの異なる機械学習モデルとして定義し（図 1

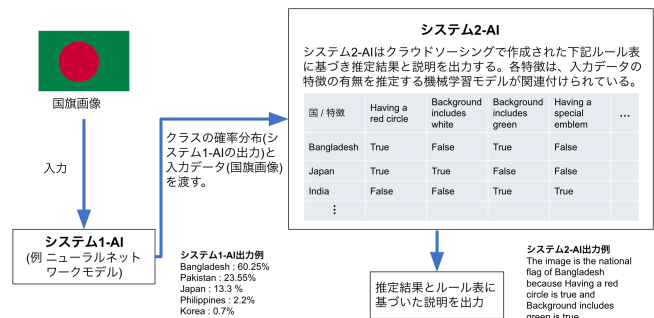


図 1: 二重過程理論の概要図を示す。入力となるデータが与えられると、そのデータは二重過程理論に基づくシステムを構成する 2 つの要素、すなわち、システム 1-AI とシステム 2-AI を通過する。システム 1-AI は、データの潜在的なクラスの確率分布を出力する任意の分類器である。システム 2-AI は、システム 1-AI の出力とシステム 2-AI のルール表を基に、論理的で人間が理解可能な説明と共に、最終的な推定結果を出力する。

参照), その内システム 2-AI の構築を行う。システム 1 は、知覚や習慣、記憶などのデータを基に素早く推定を行うシステムとして見做せるため、システム 1-AI は一般的な深層学習モデルを利用することで実現できる。一方、システム 2 は、低速であるが、論理的に思考し、推定を行うことができるシステムである。論理的に思考を行うことは、推定結果のみではなく、思考過程そのものが論理的であることを意味する。故に、決定木学習モデルのような、論理的かつ、人間が解釈しやすい構造をもった機械学習モデルを用意することで、システム 2-AI は実現できる。

問題はどのようにして、人間が理解可能な説明を出力するシステム 2-AI を構築するかである。本論文では、理解可能なシステム 2-AI を構築するために、クラウドソーシングで推定対象の各クラスに関して、人間が理解可能な特徴を収集し、これら集めた各特徴と各クラスの対応表（ルール表）を作成する。これを正

1: Eric Weiss. 'Inadequate Safety Culture' Contributed to Uber Automated Test Vehicle Crash - NTSB Calls for Federal Review Process for Automated Vehicle Testing on Public Roads. NATIONAL TRANSPORTATION SAFETY BOARD. <https://www.nts.gov/news/press-releases/Pages/NR20191119c.aspx> (参照 2020-12-17)

解を表すルール表として扱う。また、インターネットから訓練データを収集し、入力データに各特徴が含まれるか否かを推定する分類器(深層学習モデル)を特徴毎に生成する。推定時に各分類器の推定結果とルール表の一致した部分に関連付けられた特徴を推定結果の説明として利用することで、説明可能なシステム 2-AI を構築する。

我々は、システム 2-AI が正確なクラスを推定し、また人間が理解可能で、論理的な説明を出力することができるか確認するために、アジアの国旗の画像分類を題材として起用した実験 1 と実験 2 を行い、評価と今後の課題を確認した。

なお、本研究の貢献は以下の通りである。

(1) 心理学における二重過程理論が説明可能な AI の開発手法に適用できることを示唆した。

(2) ヒューマンインザループの手法を用いることで、人間が理解可能な推定基準を有した説明可能な AI を構築できることを実験的に示唆した。

2 関連研究

深層学習モデルのようなブラックボックスな AI の台頭により、機械学習モデルが出力した推定結果の説明可能性の向上は、AI の更なる普及のために解決されるべき重要な課題となっている。

既存の機械学習モデルの生成手法でも、Random Forest [5] のように、決定木学習を利用した機械学習モデルは、推定結果の根拠を示すことが可能である。しかし、そのような機械学習モデルは、推定根拠を決定木のような人間が解釈しやすい形式で示すことが可能ではあるが、自然言語のような人間が理解しやすい言葉で説明できないことや、そもそも機械学習モデルが自動生成した判断基準は、人間の判断基準と比較すると、大きな差異が有ることがあり、理解しやすいとは言い難いことがある。本研究では、クラウドワーカーが自然言語で入力した人間が理解可能な特徴を基に、システム 2-AI を構築しているため、システム 2-AI の判断基準は、人間にとって理解しやすい、人間の判断基準に近いものとなっている。

Hendricks らは Generating Visual Explanations [2] にて、AI の推定結果の説明を自然言語ベースで生成する手法を提案した。Generating Visual Explanations では、自然言語で説明を出力するため、人間にとって解釈性の高い説明を提供することができる。Generating Visual Explanations と比較すると、我々の手法では、システム 2-AI 自体がシンプルなルール表の構造をしているため、出力である説明の他に、システム 2-AI 自体も解釈性を有し、推定のプロセスを理解することが容易となっている。これは、システム 2-AI のデバッグや運用環境におけるデータトレンドの変化への対応を容易にする。

Flock [6] は、クラウドワーカーを利用し、人間が理解可能な特徴を集め、集めた特徴を基に機械学習モデルの学習を行うことで、単一の機械学習モデルや、人間の専門家を上回る性能を有した機械学習モデルを作成することができるプラットフォームを提案した研究である。我々の手法は、Flock のように、機械と人間

が得意な特徴空間の両方を一つの機械学習モデルで学習するのではなく、独立した 2 つのシステムで構成されている。これは、システムのメンテナンス性を向上させる。一方、我々の手法は、機械学習モデルが得意とする特徴空間の情報を利用できていないため、今後の手法の改良点として、決定木学習モデルをシステム 1-AI に採用し、その推定結果と判断基準をシステム 2-AI に反映することで、機械学習モデルが得意とする特徴空間の情報を利用することが挙げられる。

論理とヒューマンインザループの手法を組み合わせたもう一つの研究は、CyLog [7] である。CyLog では、論理的なルールの評価を人間が制御できるようにすることで、人間の意思決定の結果となる論理的なルールを導き出すことができるように、ゲームアスペクトの概念を導入している。我々が提案する方法が CyLog と異なるのは、論理規則がプログラマではなくクラウドワーカーによって動的に生成されることである。

3 提案手法

提案手法では、システム 2-AI の構築・運用方法について、構築フェーズと予測フェーズに分けて述べる(図 2 参照)。構築フェーズでは、クラウドワーカーから人間が理解可能な特徴を収集し、これら特徴を基にクラスと特徴の対応表(ルール表)を作成する。また、インターネットから訓練データを収集し、各特徴に対応した分類器を生成する。ルール表と分類器を組み合わせて運用することで、システム 2-AI を構築する。予測フェーズでは、入力データを受け取り、推定結果とその説明を出力する。

3.1 システム 2-AI 構築フェーズ

システム 2-AI を作成するために必要となる人間が理解可能な特徴を収集するために、各クラスに当てはまる特徴を入力してもらうタスクをクラウドソーシングで行う。クラウドワーカーには、判別対象である各クラスの人間が理解可能な特徴をテキストで入力してもらう。ここで述べている人間が理解可能な特徴とは、色や大きさ、模様などの属性を表す特徴のことを指す。なお、クラウドワーカーは、特徴をテキストで入力する際に、特定の形式に従って入力を行う必要がある(図 4 参照)。テキストのフォーマットに統一形式を採用することで、ユーザが入力したテキストから特徴のキーワードを抽出しやすくなる。

また、推定基準が重複した分類器を生成しないように、Word Mover's Distance [8] を用いて、クラウドワーカーが与えた特徴の各組み合わせについて類似度を計算し、類似度が近いグループの代表点の一つを残し、代表点に類似した他の特徴は除去する。

重複を除去して得た各特徴に対し、インターネットを介して訓練データを収集し、分類器を作成する。

分類器はそれぞれ 2 つのラベル(真か偽か)を出力するもので、入力データが、その分類器に関連づけられた特徴を有しているか否かを推定する。

また、システム 2-AI の推定結果を評価する際の比較対象として、正解として利用するルール表を作成するために、各クラスが各特徴を有しているか否かをクラウドワーカーに推定してもら

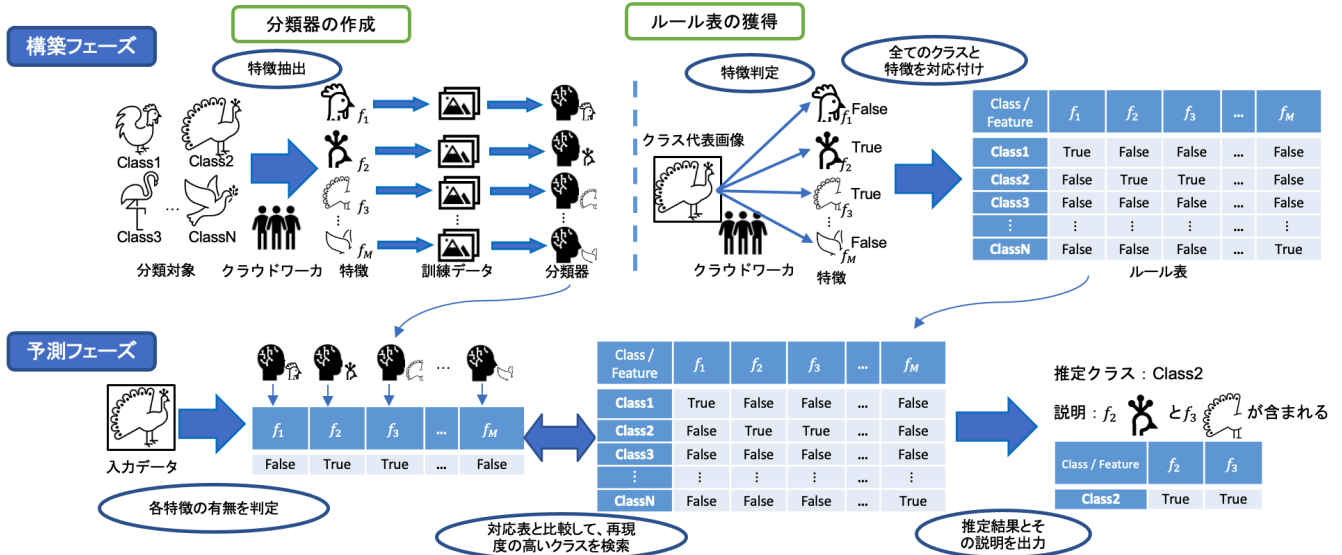


図 2: 提案手法の概要図を示す。構築フェーズでは、まず分類対象の各クラスの特徴の入力をクラウドワーカーに依頼する。得られた各特徴に関連する訓練データをインターネットから集める。収集した訓練データを用いて、入力データに各特徴が含まれるか否かを推定する分類器を作成する。また、同時にクラウドソーシングで、正解となるルール表を作成する。これらを合わせることで、システム 2-AI を構築する。予測フェーズでは、構築フェーズで作成した分類器とルール表を基に、推定結果と説明を出力する。

うタスクをクラウドソーシングで行う (図 5 参照)。このタスクでは、クラウドワーカーに各クラスの代表データの一つを示し、これに各特徴が当てはまっているか否か、真か偽の二値で答えてもらう。タスク結果として、得られた各クラスと各特徴に対応する真偽値を基に、ルール表を作成する。

そして、上記のプロセスから得た複数の分類器とルール表とを組み合わせることで、システム 2-AI を作成する。

3.2 システム 2-AI 予測フェーズ

システム 2-AI は入力データを受け取ると、各特徴に対応する分類器に受け取った入力データを渡し、返り値として各特徴が入力データに含まれるか否かの真偽値を受け取る。受け取った真偽値を、予めクラウドソーシングで用意していたルール表と比較し、最も一致したクラスを推定結果として出力する。この時、推定結果の説明として、ルール表で推定結果のクラスに含まれる特徴の中から、分類器で実際に真 (すなわち、特徴を有している) と推定されたものを、推定結果とともに出力する。ここで、ルール表と一致した特徴から、真と推定された特徴のみを出力する理由は、入力データに含まれないと推定された特徴よりも、含まれると推定された特徴の方が特徴一つあたりの情報量が多く、含まれると推定された特徴に絞ることで、得られる説明が冗長となることを防ぐことができるためである。

4 実験 1

本章では、アジア 22 カ国²の国旗の画像分類を題材として、提案手法を適用した実験 1 のワークフローと考察を述べる。

4.1 システム 2-AI 構築フェーズ

構築フェーズの概要を図 3 に示す。

人間が理解可能な特徴を集めるために、クラウドワーカーにタスク画面上で国旗の画像を 2 枚示し、右の国旗画像には含まれず、左の国旗画像にのみ含まれる特徴をテキストで、特定の形式に沿って入力してもらうタスクをクラウドソーシングで行った (図 4 参照)。なお、クラウドソーシングのプラットフォームには、Amazon Mechanical Turk³を利用した。左の国旗画像にのみ含まれる特徴を入力する目的は、クラス間で区別可能な特徴を収集することである。各国旗の特徴を得るために、2 カ国の国旗の各組み合わせを 1 タスクとして、同じタスクを 3 人のクラウドワーカーに依頼した。クラウドワーカーには 1 つのタスクに対して 0.01USD の報酬が与えられた。

クラウドソーシングの結果として、人間が理解できる特徴 (日本の国旗の場合「red circle」や「white background」など) を 2,323 個得た。

その後、意味が重複している特徴を除去し、重複のない特徴 299 個を得た。ここでは、重複した特徴を除去するために、Word Mover's Distance を用いて、収集した各特徴の組み合わせ毎に意味的距離を算出し、距離が 0.8 以下の場合、重複する特徴とみなし、意味的距離の近いグループの代表点の一つを残し、グループ内の残りの特徴を除去した。なお、意味的距離が近いとみなす閾値の設定は、開発者がデータに応じて調整する必要がある。

次に、各特徴の有無を推定する分類器の生成に必要な訓練データを収集するために、重複を除去した特徴を検索ワードとして利用し、インターネット上から、各特徴を反映した画像を収集した。ここで用いた検索ワードは、各特徴を表すテキスト (「red circle」や「white background」など) に「national flag」

2: 本論文ではアジアの国をアジアにおける日本の承認国に日本を加えたものとした。日本の承認国は総務省のウェブサイトで確認できる。
(<https://www.mofa.go.jp/mofaj/area/asia.html>)

3: Amazon が提供するクラウドソーシングプラットフォーム。(<https://www.mturk.com/>)

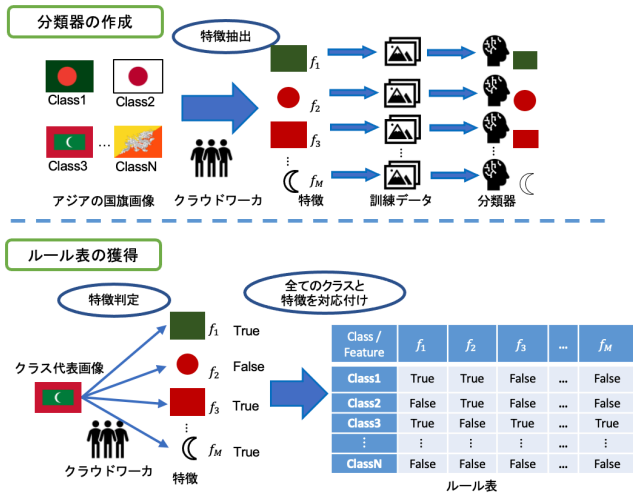


図 3: 実験 1 における構築フェーズの概要図を示す。まずアジア 22 カ国の国旗画像の特徴の入力をクラウドワーカーに依頼する。タスクの結果として得られた特徴の重複を除去し、各特徴を検索ワードとして利用することで、インターネットから訓練データを集める。収集した訓練データを用いて、入力データに各特徴が含まれるか否かを推定する分類器を作成する。また、同時にクラウドソーシングで、正解となるルール表を作成する。これらを組み合わせることで、システム 2-AI を構築する。

というテキストを追加したものである。例としては「red circle national flag」が挙げられる。「national flag」を特徴を表すテキストの後ろに追加した理由としては、システム 2-AI の推定対象である国旗画像の特徴をよく反映した訓練データを収集するためである。

集めた画像を RGB カラーモードを保持したまま 200 × 200 の画像に整形し、これらを訓練データとして、関連する特徴の有無を推定する分類器を複数生成した。学習過程では、正例は画像検索で収集した各特徴をよく反映した画像を利用し、負例は他の特徴の画像を特徴毎に 1 枚ずつ抽出したものを利用した。

また、各クラスが、各特徴を含んでいるか否かを問うタスクのクラウドソーシング（図 5 参照）を行った。なお、このタスクは、同一のタスクを 3 名のクラウドワーカーに割り振り、1 度のタスクで 13 個の特徴の有無をクラウドワーカーに判定してもらった。クラウドワーカーには、報酬として 0.03USD が支払われた。このクラウドソーシングから得られた回答からクラスと特徴のルール表を作成し、正解を表すルール表として扱った。

Instructions:

Please write as many features of the left national flag that the right national flag does not have as possible. The answer should be in one of the following forms:

-Form1-
There are/is < a number > [optional < color >] < object > in the left flag.
-Example1-
There is a red circle in the left flag.

-Form2-
The background colors include < colors > in the left flag
-Example2-
The background colors include white in the left flag





Your answer.

Your answer

Add answer Remove

Submit

図 4: 各クラスが有する特徴の収集のために行うクラウドソーシングタスクのデザインを示す。タスクの下部にはクラウドワーカーが見出した人間が理解できる特徴をテキストで入力するフィールドがある。クラウドワーカーは入力フィールドを利用して、右の国旗画像には含まれず、左の国旗画像には含まれる特徴を入力する。これは、クラス間の判別に有用な特徴を集めるためである。



Instructions:

Would you please judge whether each of following features matches the national flags above. If so, please switch the radio button to "True"

There is/are no any division in the flag ☒ False ☐ True

There is/are no yellow color present in the flag ☒ False ☐ True

There is/are no pointy symbols in the flag ☒ False ☐ True

There is/are triangle like angles in the flag ☒ False ☐ True

There is/are a red circle in the center in the flag ☒ False ☐ True

There is/are a circle shape in the flag ☒ False ☐ True

There is/are no horizontal striped shaped present in the flag ☒ False ☐ True

There is/are a big red circle in the flag ☒ False ☐ True

The background colors include orange-red in the flag ☒ False ☐ True

There is/are A DRAGON in the flag ☒ False ☐ True

The background colors include YELLOW in the flag ☒ False ☐ True

There is/are a dragon in the flag ☒ False ☐ True

There is/are a black and white dragon in the flag ☒ False ☐ True

Submit

図 5: ルール表の作成に必要となる、クラスと特徴の対応関係の判定を依頼するクラウドソーシングタスクのデザインを示す。クラウドワーカーに各クラスの代表となる画像データを示し、タスク下部で示された各特徴が画像データに含まれるか否かを判定した結果を入力してもらう。

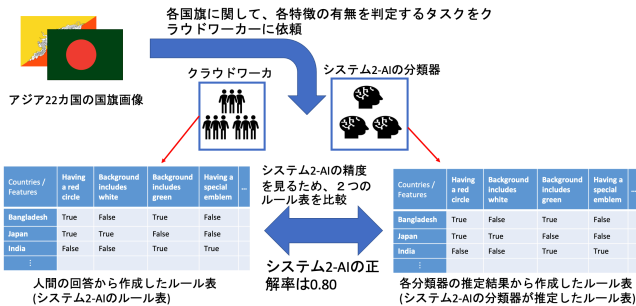


図 6: 実験 1 における予測フェーズの概要図を示す。構築フェーズで、クラウドワーカーに各国旗に各特徴が含まれるか否かの判定を依頼し、作成した正解となるルール表と、各分類器の推定結果を基に作成したルール表とを比較し、モデルの精度を算出する。

4.2 システム 2-AI 予測フェーズ

予測フェーズの概要を図 6 に示す。

システム 2-AI の評価として、人間の回答を基に作成したルール表と、各クラスの代表画像に対するシステム 2-AI の推定結果を基に作成したルール表との比較を行った。図 7 にその結果を混同行列として示す。

4.3 考 察

人間の回答を基に作成したルール表と、各クラスの代表画像に対するシステム 2-AI の推定結果を基に作成したルール表との比較を行った結果 (図 7 参照)、正確率:0.80、適合率:0.192、再現率:0.224、特異率:0.876、F 値:0.207 となった。正確率はある程度高いが、適合率、再現率ともに低い値であった。

正確率が高く現れている理由としては、特徴を収集するクラウドソーシング時に、各クラスに特有の特徴を集めるようにタスクデザインを行ったため、特徴全体に比べ、各クラスに当てはまる特徴の割合が小さいことが原因であると推測できる。すなわち、人間が判定した特徴のほとんどは False(Negative) であり、これが正確率を高める要因となっている。

この結果で重要視すべきは、適合率、再現率であり、両方の値が低い値を示していることから、入力データに存在しているはずの特徴を正しく検出できていないことである。この原因としては、各特徴に関連付けられた分類器の性能が悪く、分類器の学習に用いた訓練データが適していなかったことが原因であると推論できる。

各特徴に関連付けられた分類器の性能を向上させるためには、今回の手法のように、各特徴を反映した画像を収集するのではなく、各クラスの画像を収集し、各特徴に関連付けられた分類器の訓練時に、クラスの画像と、その特徴がクラスの画像に含まれているか否かのラベルとを入力として与えることで、精度の向上を図ることが方法として挙げられる。

5 実 験 2

本章では、実験 1 の考察を踏まえ、各クラスを表す画像を訓練データとして用いることが、特徴に関連付けられた分類器の性能の向上に有効であるか検証を行った実験 2 について、その

対象	個数
収集した特徴	2323
重複除去後の特徴	299

分類器による予測結果

	特徴あり	特徴なし	小計
人間による予測結果			
特徴あり	172	597	769
特徴なし	720	5,089	5,809
小計	892	5,686	6,578

図 7: 各クラスが 299 個の特徴それぞれを含んでいるか否かの判定をクラウドワーカーに依頼し、得られた回答結果を基に作成したルール表と、299 個の特徴それぞれに関連付けられた分類器に、各クラスの代表となる画像を入力し、得られた推定結果を基に作成したルール表とを比較した結果を示す。

ワークフローと考察を述べる。

5.1 システム 2-AI 構築フェーズ

実験 2 における構築フェーズでは、実験 1 で収集し、重複除去を行った特徴とルール表を利用し、実験 1 とは異なる方法で各特徴に関連付けられる分類器の学習を行った。

実験 1 では各特徴を表す画像 (赤丸を表す画像など) をインターネットから収集し、分類器の訓練データとしたが、実験 2 では、各クラスを表す画像 (日本の国旗画像など) をインターネットから収集し、訓練データとして利用することで、分類器の学習を行った。

各特徴に対応付けた分類器の学習に利用する訓練データとして、ルール表を基に、正例となるデータとしては、特徴を有すると人間によって判定されたクラスの画像を、負例となるデータとしては、特徴を持たないと判定されたクラスの画像を用いた (図 8 参照)。なお、正例となるデータと負例となるデータの数は同程度になるようにした。これら訓練データを利用して、分類器を生成した。

5.2 システム 2-AI 予測フェーズ

実験 1 と同様に、システム 2-AI の評価として、人間の回答を基に作成したルール表と、各クラスの代表画像に対するシステム 2-AI の推定結果を基に作成したルール表との比較を行った。図 9 にその結果を示す。

実際にシステム 2-AI に国旗の画像を与え、299 個の分類器それぞれで関連する特徴の有無を推定し、得られた結果を人間の回答を基に作成したルール表と比較した。比較の結果、再現率の高いクラスから順にランキング形式で推定結果と推定結果の説明を出力した。この出力結果の例を図 10 に示し、各国旗の代表画像全体を評価した時、正解のクラスがどの順位に現れたかを図 11 に示した。なお、推定結果の説明としては、人間の回答を基に作成したルール表で真 (特徴を有する) と判定されており、かつ分類器でも真と推定された特徴を説明に起用した。また、再現率をランキングに利用した理由としては、クラスによって含有する特徴の個数にばらつきがあり、このばらつきを考慮するためである。

システム 2-AI の推定結果の説明の正確性の評価を行うため

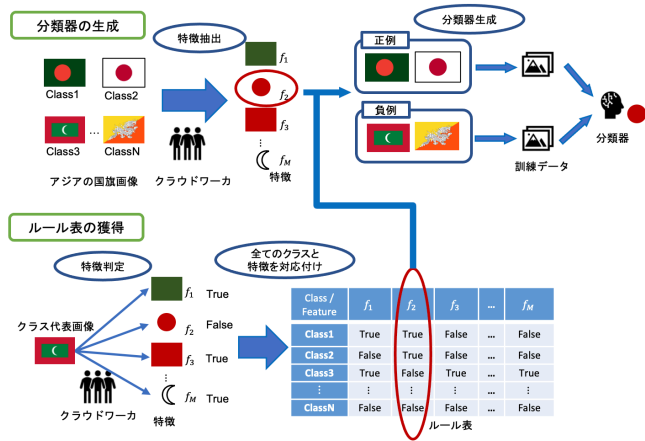


図 8: 実験 2 における構築フェーズの概要図を示す。アジア 22 カ国の国旗画像の特徴の入力をクラウドワーカーに依頼する。タスクの結果として得られた特徴の重複を除去し、これら特徴を基にクラウドソーシングで特徴とクラスの対応表 (ルール表) を作成する。ルール表を利用し、インターネットから収集した各国旗クラスを表す画像を、正例と負例に分け学習し、特徴毎に分類器を生成する。分類器とルール表を組み合わせることで、システム 2-AI を構築する。

に、クラウドワーカーの回答を基に作成したルール表が正確か否かの検証と、推定結果の説明に含まれる特徴が正確か否かの検証を行った。クラウドワーカーの回答によるルール表の検証方法としては、執筆者と研究室内で募集した実験協力者 1 人の計 2 人が、それぞれルール表を作成し、この 2 つのルール表の論理積を取ることで、1 つのルール表としたものを、クラウドワーカーの回答を基に作成したルール表と比較することで行った。また、推定結果の説明の検証方法としては、クラウドワーカーの回答を基に作成したルール表と分類器の両方に真と推定された特徴が推定結果の説明の構成要素となっており、これらの特徴の中で執筆者と研究室の実験協力者の 2 人の回答を基に作成したルール表においても真と推定されたものを正確な特徴としてみなし、正確な特徴が説明に含まれる特徴に占める割合を調べることで評価を行なった。

また、推定結果の説明の冗長性の評価としては、クラウドワーカーの回答によるルール表と、執筆者と研究室の実験協力者の回答によるルール表の論理積を取り作成したルール表から、各クラスに特有な特徴の組み合わせを取得し、実際の推定結果の説明に含まれる特徴が、そのクラスを推定するのに必要十分か検証することで行なった。

5.3 考察

5.3.1 システム 2-AI の性能評価

人間の回答を基に作成したルール表と、各クラスの代表画像に対するシステム 2-AI の推定結果を基に作成したルール表との比較を行った結果 (図 9 参照)、正解率:0.983、適合率:0.894、再現率:0.965、特異率:0.985、F 値:0.928 となった。実験 1 とは異なり、全ての指標が高い値を示した。

しかし、これら算出した指標の内、適合率のみが他の指標の値よりも、少し低い値を示した。この理由としては、実験 1 では、

分類器による予測結果			
	特徴あり	特徴なし	小計
人間による予測結果			
特徴あり	742	27	769
特徴なし	88	5,721	5,809
小計	830	5,748	6,578

図 9: クラウドワーカーの回答結果を基に作成したルール表と、各クラスの代表画像に対する各分類器の推定結果を基に作成したルール表とを比較した結果を示す。

特徴を表す画像を訓練データとして利用していたが、実験 2 では、特徴を有すると人間によって判定されたクラスの画像を訓練データとして利用したため、画像の特徴部分のみを正しく学習したのではなく、特徴を有すると人間によって判定された複数のクラスの画像間で共通する特徴以外の部分も学習し、推定を行ったことが一因であると推察できる。

また、推定結果 (図 10, 11 参照) に関しては、概ね正しい結果が出力された。なお、推定結果の定量的評価として、ランキングの評価指標として利用される MRR を算出すると、0.977 を示した。なお、MRR は順位の逆数の平均をとったもので、0 から 1 の値を取り、より 1 に近いほど、正確に推定ができていることとなる。MRR の値 0.977 は、平均的に、ランキング 1 位に正しいクラスが位置していることを示している。ゆえに、推定結果全体として見るならば、まずまず正しい推定ができています。

以上のことから、各特徴に関連付けられる分類器の学習に、特徴を有すると人間によって判定されたクラスの画像を訓練データとして利用することで、システム 2-AI の精度の向上を見込めることが示唆された。

しかし、前述したように、画像の特徴部分のみを正しく学習したわけではなく、特徴を有すると人間によって判定された複数のクラスの画像間で共通する特徴以外の部分も学習したと推測できるため、今後の課題として、運用環境での入力データのトレンドの変化など、未知のクラスの入力データが与えられた場合に対し、正しく特徴を認識できるか検証することや、LIME [9] などの手法で、画像のどの部分が大きく推定結果に影響しているかを見て、正しく特徴を学習できているか検証することが挙げられる。

5.3.2 推定結果の説明の評価

推定結果の説明の評価として、理想的な説明の条件として挙げた、正確性、非冗長性、人間による理解可能性の 3 つの観点から評価を行なった。正確性に関しては、クラウドワーカーの回答から作成したルール表と執筆者と研究室内で募集した実験協力者の 2 人から作成したルール表を比較することで行なった。結果としては、正解率:0.90、適合率:0.39、再現率:0.67、特異率:0.92、F 値:0.50 となった。再現率が 0.67 であり、7 割近くの正確な特徴をシステム 2-AI が認識できていることになるが、一方、適合率が 0.39 と低く、システム 2-AI の性能評価でも述べたように、これは特徴部分だけではなく、その特徴を含むクラス全体を表した画像を訓練データの正例として使ったため、特徴部分以外も学習してしまったことが影響していると推察できる。

また、説明の正確さの検証を、クラウドワーカーと分類器、そし




順位	予測	説明
1	 Maldives	RED, dark red, a white moon symbol, a semi moon, A CRESCENT MOON, a white crescent moon, a green stripe, a red stripe, maroon, three colors present, raspberry red, a white design, a white moon's picture, ...
2	 Laos	RED, dark red, a red stripe, a white moon's picture, a moon, three colors, a white moon, white, red
3	 Indonesia	RED, a red stripe, raspberry red, a white design, color combinations, white, red

図 10: システム 2-AI にモルディブ (Maldives) の国旗画像を入力データとして与えた時の推定結果と、推定結果の説明を示す。なお、推定結果のランキングは、入力データに対する推定結果を再現率が高いものから順に並べたものである。

て執筆者と実験協力者によって真と推定された特徴を正しい特徴として見なして行なった。その結果、図 12 に赤字(下線部)で表した特徴が正しい特徴であり、図で例として挙げているモルディブにおいては、説明に用いられる特徴の内、正しい特徴の割合は、0.48 となった。また、赤丸で囲んだ特徴のように、「while moon's picture」の「while」は「white」の誤字であり、また、「a red stripe」のように、縞模様を表す特徴が、縞模様の無いモルディブの国旗に対し、検出されてしまっているなど、説明として不適切な特徴が含まれてしまっていた。これを改善するには、クラウドワーカーから特徴を収集する段階や、クラウドワーカーにクラスと特徴の対応を推定してもらう段階で、クラウドワーカーの回答を、他のクラウドワーカーに検証してもらう工程を取り入れるなどの対策が必要である。

説明の冗長性の検証を、クラウドワーカーの回答を基に作成したルール表を基に、各クラスに特有の特徴の組み合わせを見つけることで行なった。その結果、図 13 の赤丸で囲まれた特徴のみ推定できていれば、モルディブであると特定することができ、モルディブの説明は冗長であることが判明した。一方で、ラオスの説明で利用されている特徴の全ては、クラウドワーカーの回答を基に作成したルール表では、モルディブにも当てはまってしまうため、ラオスとして推定するには説明が不十分である結果になった。これを改善するには、クラスに特有な特徴の組み合わせの中から、推定結果に一致するものを起用すれば、説明を非冗長にすることができると推察できる。

説明の人間による理解可能性については、クラウドワーカーが各国旗を見て認識した特徴を利用しているため、各特徴に関しては、理解可能であると推察できる。しかし、説明が冗長であり、入力画像が含むと推定された特徴の列挙になってしまっているため、説明の要点を得にくくなっている。ゆえに、改善策としては、冗長な特徴を減らし、説明として必要十分かの評価も説明に含めることで、理解可能性の向上を図ることなどが挙げられる。

6 まとめと今後の課題

実験 1 の結果から、各特徴に関連付けられた分類器の訓練が適切ではなかったこと、その一方で、実際に推定結果とその説明

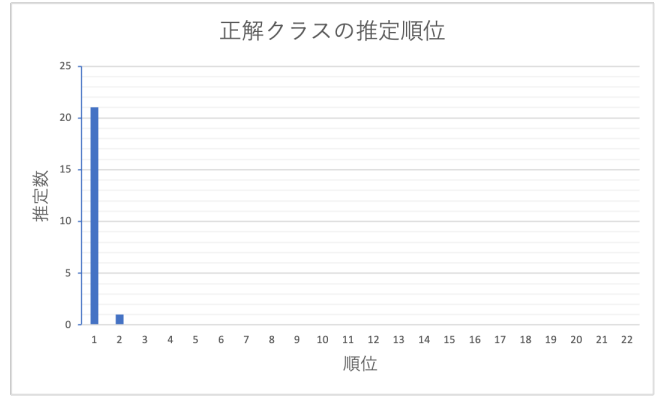


図 11: システム 2-AI に各国旗の代表となる画像を入力として渡し、推定を行った結果、本来推定されるべき正しいクラスが、ランキングのどの順位に現れたかをグラフで示す。

順位	予測	正解率	説明
1	 Maldives	0.48	RED, dark red, a white moon symbol, a semi moon, A CRESCENT MOON, a white crescent moon, a green stripe, a red stripe, maroon, three colors present, raspberry red, a white design, a white moon's picture, ...
2	 Laos	0.44	RED, dark red, a red stripe, a white moon's picture, a moon, three colors, a white moon, white, red

図 12: 赤字(下線部)は、説明に用いられる特徴の中で、クラウドワーカーと分類器、そして執筆者と実験協力者の 2 人に、そのクラスが有する特徴として推定されたものである。正解率は、説明に含まれる特徴数に占める、正確な特徴の割合である。また、赤丸で囲まれた特徴は、誤字や、本来含まれない特徴であり、不適切と見做せる例を挙げたものである。



順位	予測	説明
1	 Maldives	RED, dark red, a white moon symbol, a semi moon, A CRESCENT MOON, a white crescent moon, a green stripe, a red stripe, maroon, three colors present, raspberry red, a white design, a white moon's picture, ...
2	 Laos	RED, dark red, a red stripe, a white moon's picture, a moon, three colors, a white moon, white, red

図 13: モルディブの推定結果の説明の内、赤丸で囲まれた特徴のみが推定できれば、他の特徴は推定できなくても、モルディブであると推定できる。一方、ラオスの推定結果の説明に現れる特徴の全ては、クラウドワーカーにモルディブに含まれると推定されている特徴でもあるため、ラオスと推定するには、説明として不十分となっている。

を出力すると、半数程度は正しいことが判明した。我々の提案手法は、正しい推定結果と、推定結果の説明を出力できる可能性があるかと推察できるが、これと同時に、個々の分類器の精度という大きな問題があることも明らかとなった。

そこで、実験 2 では、分類器の学習で利用するデータを各特徴ではなく、各クラスに関係するデータを用いることで精度の改善を図った。これにより、分類器の精度は高くなったが、分類器が関係する特徴以外の部分を見て推定を行っている可能性もある。ゆえに、LIME [9] などの手法で特徴のみを正確に学習できているか検証を行うことや、Multiple instance 学習 [10] を利用する

ことで、画像内の特徴の領域をアノテーションした訓練データを用意することなく、特徴のみを自動的に学習させることなどが、今後の改善策として挙げられる。

他の課題として特徴の重複の除去方法が挙げられる。提案手法では Word Mover's Distance を利用し、特徴間の意味的距離が近いものを重複と見なし除去していたため、開発者が意味的距離について、重複とみなす距離の閾値を設定する必要があった。一方、特徴の重複除去に、Word2Vec [11] で特徴を表すテキストデータをベクトル化した後、クラスタリングを行い、各クラスターの代表点を取得する方法を利用することで、開発者の主観の介入を防ぎ、また、重複除去後の特徴数を設定できるようになる。

本研究の次のステップとしては、上記の課題の改善に加え、システム 1-AI とシステム 2-AI の連携が挙げられる。正解となるルール表から、各クラス特有の特徴または特徴の集合を予め検出しておき、システム 1-AI の推定したクラスをシステム 2-AI が受け取り次第すぐに、そのクラス特有の特徴に関連付けられた分類器で推定を行うことで、システム 2-AI 単体で推定を行うよりも、時間コストを大幅に減少させることができると推察できる。

また、手法の評価に関しては、本論文では、システム 2-AI の全体としての精度を利用したが、その他にも、本研究では多くの分類器を利用し推定を行うため、計算コストや時間コストを考慮した評価も取り入れる必要がある。以上のことが今後の課題として挙げられる。

謝 辞

本研究の一部は、JST CREST JPMJCR16E3 と JST 未来社会創造事業 JPMJMI19G8 の支援を受けたものである。ここに謝意を示す。

文 献

- [1] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML2017)*, pp. 1885–1894, 2017.
- [2] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *Computer Vision (ECCV2016)*, pp. 3–19, 2016.
- [3] Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pp. 5540–5552, 2020.
- [4] Daniel Kahneman. *Thinking, FAST AND SLOW*. Farrar, Straus and Giroux, 2011.
- [5] Leo Breiman. Random forests. *Machine Learning*, Vol. 45, No. 1, pp. 5–32, 2001.
- [6] Justin Cheng and Michael S. Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW2015)*, pp. 600–611, 2015.
- [7] Morishima Atsuyuki, Fukusumi Shun, and Kitagawa Hiroyuki. Cylog/game aspect: An approach to separation of concerns in crowd-sourced data management. *Information Systems*, Vol. 62, pp. 170–184, 2016.
- [8] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Wein-

berger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference in Machine Learning (ICML2015)*, pp. 957–966, 2015.

- [9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2016)*, p. 1135–1144, 2016.
- [10] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, Vol. 89, No. 1, pp. 31–71, 1997.
- [11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR2013)*, 2013.