技術論文

人工知能の説明性と不確かさの統一理解に向けて

仲村　佳悟
Keigo Nakamura

概　要

AIの信頼性を確保するために説明可能性AI(XAI)の手法が数多く提案されてきたが未だ万能なXAIは存在していない.説明可能なAIを様々な分野でより効果的に適用するためには「説明」自体の性質を理解することが重要である.本稿では信頼性が重要な品質管理と意思決定理論の考え方に従ってAIにおける説明可能性と不確実性の関係を提案した.またAIモデルを用いて新たな不確かさの導出法とこの関係性の検証方法を考案した.

1.　はじめに

AIの発展が進むにつれて説明可能AI(XAI)の重要性が指摘されており,様々な手法が開発されている.一般的にAIの分野では「AIがなぜその出力をしたのかを人間が理解できるように説明できること」がAIの説明性として“定義”されており,その定義に従って研究が進められてきた[2].XAIの標準的手法であるLIME[3]やGrad-CAM[4]は予測に影響を与えた要素や領域を示すことでその出力の根拠としている.

一方で実用上ではただ重要な要素を指示されただけでは適切な説明にはならない場合が多い.例えば不良品検査において傷の箇所を示すことは既存のXAIでも可能だが,AIの説明が無くても傷がついているのは見ればわかってしまう.本当に知りたいのは「なぜ,どの工程で傷がついたのか」や「他に傷を見逃していないのか？」の根拠であったりする.

このようにXAIから出力される情報と実際のニーズが一致していないため,XAIが適切に活用されているとは言えない.そこで「本当に必要な説明」はどのような特性を持っているのかを考えることでより便利なXAIが作れないかというのが本研究のモチベーションである.

2.　「説明」とはどういう情報だろうか？

AIの説明性は説明によって「AIの信頼性」を担保するために必要であるとして指摘されている[2].AIに限らず信頼性が重要である分野ではすでに信頼性を向上させるための多くの知見や説明が蓄積されており,AIをこれらの分野に適応する際には,当該分野の信頼性に関する知見とAIの説明性の性質が合致していない限り,十分な説明になり得ない.そこでアイシンとしても重要な「品質管理」と「意思決定」において説明となる情報がどのような性質を持っているのかを調べた.

2.1　品質管理において必要な情報

アイシンのような製造業でのAIの適応先として良品検査があげられるが,一つの不良品でもお客様に迷惑をかけるため,AIの判断においても高い信頼性が求められる.品質管理(QC)における信頼性とはJISによると「アイテムが与えられた条件で規定の期間中,要求された機能を果たすことができる性質」と定義されている[5].QCでは信頼性の指標として測定の「バラつき」を調べ,バラつきの原因を特定し対策を行うことで低減し品質を高める活動を行っている[6].この文脈において,品質管理における「信頼性」とはバラつきの大きさであり,バラつきを削減する「情報」が信頼性を高めるために必要な情報と言える.

2.2　意思決定において必要な情報

医療,自動運転,経営判断等の意思決定を伴う分野におけるAIの適応においても信頼性が重要である.人は認識能力の限界から常に「不確実な情報の下」で意思決定を行っており[7],より適切な判断を行うために様々な“確実な情報”を収集している[8),9].例として「A社に融資するかしないか」という判断を考えてみると,計画の出来や実現性,経営者の経歴や信用情報等,様々な情報を得ることで不確実な情報が減り,判断の「信頼性」を高くできる.

意思決定の分野では測定可能な不確かさを「リスク」とよび,想定すら難しいものを「Uncertainty=不確実性」と呼ぶが,意思決定においても信頼性を上げるため

に必要な情報は「リスクを減らす情報」と言える.

2.3　説明性と不確かさ

今の説明は各専門分野に存在している厳密な言葉の定義に基づいたものではないため専門家の方々から批判があるかもしれないが,定性的には「不確かさやリスク」が信頼度の尺度として存在し,それを下げる情報を得て不確かさを下げることで信頼度を上げていると言える.つまり,これら分野において必要な説明とは「不確かさを下げる情報」のことと言える.

AIの説明性においても同様の性質が必要である.つまり今現在どれくらい確実な予測になっているかの「不確かさ」の導出であり,それが「付加的な情報＝説明」によってどのように変化するのかというのが説明性の本質だと言える.

LIME等において重要な要素を示すことが説明になり得るのはその要素が意思決定の確度を高める情報であるということを“予め”知っているからに他ならない.先の融資の例で休日前は担当者の機嫌がよく,金曜日だと融資されやすい場合「曜日」も非常に重要な要素になるが,LIMEが「曜日」を示してもリスクが下がるのか不明瞭なため一般的には説明になり得なくなる.

また,“より信頼性の高い”予測を求める場合にはLIME等の様に出力の説明をするだけでは状況によっては不十分であることもわかる.これらの手法は既に学習されたモデルから説明を導出するので,説明を導出しても出力そのものは何も変化もしない.モデルが90%で予測したものが説明によって99%にはならないし不確かさも変化しない.説明によって変わるのは人間側の「AIの予測を信頼するかどうか」のリスクであって出力そのものではない.そのためAI予測の信頼性を上げるには「付加的な情報＝説明」によって出力の不確かさを適切に削減することができるものでなければならないと言える.

3.　説明をAIモデルに落とし込む

ここまでの議論は(AI分野では調べた限り指摘されていないものの)既に知られた概念であって,新しい発見は特にない(本研究は1978年にノーベル経済学賞を受賞したサイモン，2002年に受賞したカーネマンが提唱した概念を元にしている).「AIの研究」としては何らかの「説明可能モデル」を構築し,AIにおける不確かさを導出することで,その説明による不確かさの変化を定量的に評価する必要がある.前述の様にLIME等の出力結果を説明する手法では人間の信頼度が変わるだけで定量的な測定ができないため,不確かさと説明性双方が定量的に導出できるモデルを構築することが必要である.

例として2つのAIモデルを考えてみる.図1は説明を予測したのちにその説明の寄与によってタスクを予測するConcept Bottleneck model(CBM)と呼ばれるモデルで[10],図2は説明とタスクを並列に同時予測するMulti-Task Model (MTM)というモデルである.人間の考え方として自然なのはCBMな気がするが,双方とも説明とタスクを予測することができ,予測精度はMTMの方が高い.そのため単純に精度で比較をした際にはMTMに軍配が上がってしまう.一方で「不確かさ」に着目するとCBMの方は不確かさが誤差伝搬の要領でタスクに変化を与える可能性があるのに対して,MTMでは説明の不確かさがタスクに影響を与えることはない.そのため2つのモデルを比較することで不確かさと説明性の関係を調べることができ,かつ“説明性の高さ”の定量的な指標としても不確かさを活用できる可能性がある.

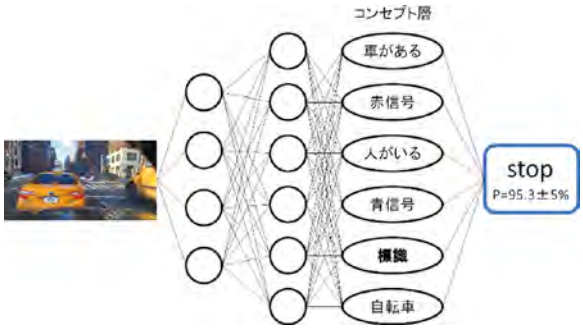

図1　CBMの概要図

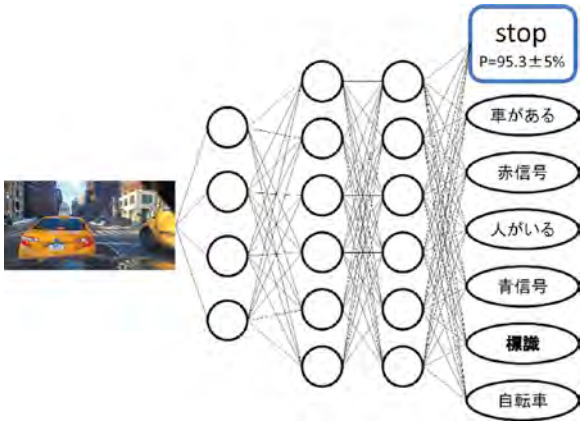

図2　マルチタスクモデルの概要図

4.　AI予測の不確かさを求める

AIの不確かさの導出方法も数多く提案されているにも関わらず,本研究に適応可能な手法は先行研究にはなく新たに考案する必要があった.図3の左図は代表的な手法であるDeep Ensemble法(DE)で導出された不確かさだが[11],DEで導出される不確かさは「分類曲線のばらつき」であり,分類そのもののばらつきを表しているわけではない(他の手法も同様の不確かさを与える).

分類問題においてより「尤もらしい」不確かさを導出できるのはガウス過程である.ガウス過程では不確かさを

“クラスタ”との距離として計算する(図4).この距離についての関数をRadial Basis Function(RBF)と言い,以下の様に定義できる.

$$RBF = exp\left(-\frac{|x_{test} - x_i|^2}{2\sigma_i{}^2}\right)$$

x_{test}はテスト画像の特徴量,x_iは各ラベルの特徴量の平均値,σ_iは各ラベルの分散である.

深層学習においてガウス過程と同様の不確かさを導出する手法としてdeterministic uncertainty quantification(DUQ)[12]が提案されているが,説明性と同時に扱うにはロス関数の計算が少々煩雑である.そこで,説明可能モデルとの組み込みやすさを考えてより簡潔に導出する手法を開発した.

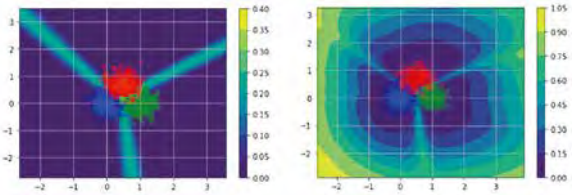

図3　クラスタ分類の不確かさ.左:Deep Ensemble法,右:本研究手法

4.1　特徴量空間にクラスタを作る

RBFは同じ特徴を持つものが特徴量空間上で“クラスタ”になっていれば適応可能である.そこで,クラスタが作成できる学習手法として全く別の文脈において広く使われている学習である対照学習をベースにしたクラスタ化を行った[13].対照学習は主に半教師あり学習で使われるもので以下のようなロス関数を用いる.

$$\mathcal{L}_{cont} = \sum_i \frac{-1}{|P(i)|} \sum_p \log \frac{\exp(c_i \cdot c_p/\tau)}{\sum_a \exp(c_i \cdot c_a/\tau)}$$

P_iはバッチ内のiとは異なるpositiveの数の割合,c_aとc_pはそれぞれanchorとpositiveの特徴量,τは温度パラメータである.このロスは「同じものを近づけて,異なるラベルを遠ざける」様に学習を行っているため,同じラベルの特徴量ベクトルが距離的に近くに配置され,クラスタ化しているはずである.

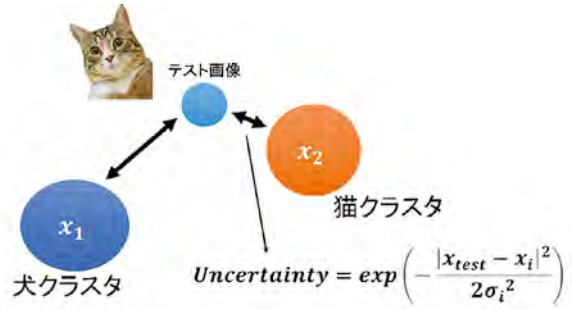

図4　DUQの不確かさの概要

簡単な実験としてCIFAR10データセットとResnet50を用いて学習を行い,不確かさの評価を行った.

4.2　不確かさの評価

不確かさの評価としてはAUROC(Area Under ROC)を用いた.他の手法と比較した結果が表1である.DE,DUQの他に通常の分類学習のロスであるCross Entropy Loss(CE) ともう一つの主要な不確かさの導出手法であるMonte Carlo Dropout(DO)[14]と比較を行った.本研究手法は他手法と同程度の精度を保ちつつ,AUROCも高い指標を示している.また,図3の右図は本研究手法が予測する不確かさであり,ガウス過程の様にデータの存在しない部分で不確かさが高くなっているのがわかる.その他エントロピーの積算分布の評価においても本研究手法がよいことを確認した.

表1　精度とAUROCの比較

model	Accuracy	AUROC
CE	94.1	0.897
DE	93.5	0.885
DO	93.5	0.895
DUQ	86.9	0.909
本研究手法	95.3	0.938

テストデータを入力した時の結果を図5に示す.テスト画像がクラスタ内にある場合($|x_i - x_{test}| \sim \sigma$),RBFは$e^{-1}$~0.6程度になるので,RBFの値が“自信度”として解釈可能になる.既存モデル(DUQ)ではデータに含まれていない分類の画像(Obi-One)を入力した際に間違った予測を高いスコアで予想してしまっているのに対して,本研究手法では全クラスのスコアが同程度出力されており「わからないデータである」ということを示すことができている.

最も不確かさの指標にも懸念点がある.分類の不確かさの指標として直観的なのはエントロピーである(図5の様に等確率の時最も大きくなり,一つのラベルのみ出力されるとき最も小さくなる).分類問題で最も一般的に使われる交差エントロピーロス(CE)はエントロピーを「最小化」する学習のため最も不確かさが少ないモデルになる.ある意味で評価するべき指標そのものが最適化されてしまっている状況のため,その他の評価指標を使わざるを得ず,評価を難しくしている.適切な評価指標については今後の研究が待たれる部分である.

5.　今後の方針

不確かさを導出したのでそれが説明によって削減されることを示したい.その概要を図6に示す.図の様にタスク分類のクラスタ内に説明のクラスタが内包されているような状況を考えると「Aの説明ならばA」という十分条件な論理を集合論的に示すことができる.このような配置を特徴量空間に実現することができれば“論理構

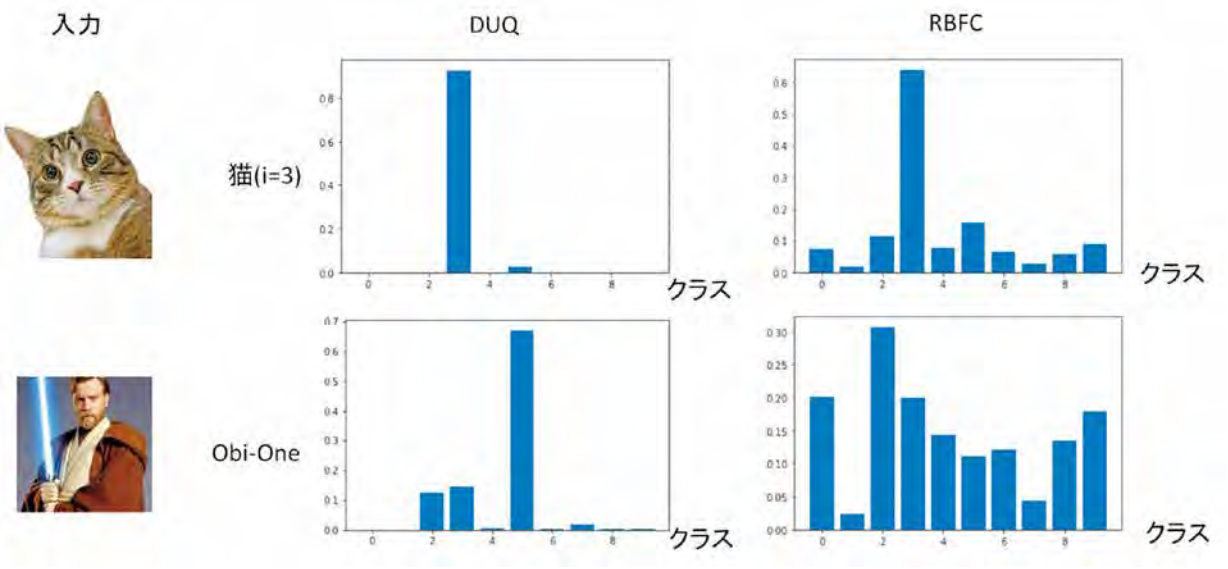

図5　DUQと本研究手法のIn-distribution とOut of distribution データを入れた際の不確かさ

造”を自然にAIに導入することができる.

また,不確かさの増減については以下の様に考えることができる.あるタスクが予測された時($|x_i-x_{test}|\sim\sigma$),RBFを用いた不確かさは“クラスタの大きさ”程度になる.説明がない場合の不確かさはタスクのクラスタの大きさのままになるが,説明が選ばれた場合はその不確かさは説明の不確かさと同じ程度になるはずである.そのため説明が選ばれるとタスクのクラスタの大きさよりも小さくなるはずであり,説明によって不確かさが削減できることを示せる.このような配置の作り方については不確かさと同様に対照学習を用いるアイディアを考案している[15].

一方で全ての説明ができるわけではない.説明は十分条件だけでなく必要条件や否定が入る場合もある.また,論理をつなげていくほど,特徴量空間の範囲を狭めていく＝不確かさを下げていくことができるが,複雑な構造を意図的に特徴量空間を作っていくことは恐らく困難である.また,今回は「AIの予測の不確かさ」で計算したが,測定においても様々な原因から「不確かさ」が入り込むように「何の不確かさ」を削減したいのかにも依存する.LIMEの例の様に「人間の信頼度」のような数値化できないものや,トロッコ問題の様に人間の思考においても解決できないリスクも存在する.

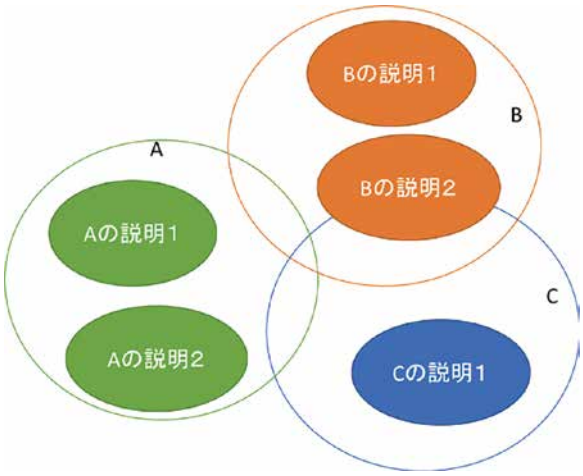

図6　説明を加えた特徴量空間の概要図

そもそもAIが不確かさを元に“適切な”予測をしているかどうかの評価には何らかの参照が必要であるが,「不確かな情報を元にした意思決定」に関してはDempster-Shafer理論[17),18)]やプロスペクト理論[16)]等が考案されているものの未だ定式化できていないため,何が適切かの定量的な評価は困難を極めると思われる.

6. おわりに

「不確かさの軽減によって信頼性を高める」という考え方は前述の様にAI技術とは関係ない一般的な考え方である.2024年にヨーロッパではEU規制法[19)]が制定されたが,高リスクな分野へのAI利用の際には予めリスクアセスメントを行い,許容できるリスクの範囲内に収まるようにAIへの要件を決める法規制が出来つつある.この場合AIの信頼性は規制の要件によって担保されるので,要件を満たすAIが作られさえすれば説明性が無くても問題はない.また,社会的なAI受容によって「AIを使う

こと」のリスクが解消されてきた現在では「AIがなぜその出力をしたのかを人間が理解できるように説明できる」ことを殊更強調する必要がなくなり,現状レベルの説明性の研究については大方役目を終えたような印象である.

一方で完全自動運転のような「AI自身が意思決定を行う」アプリケーションの場合には本研究で示唆したようにAIが適切な情報を元に判断を行っていることを示すだけでなくそれによって判断の信頼性を改善していく必要があるためさらなる研究が必要になる.しかし同時にAIに責任ある行為を担わせる倫理的な問題や責任の所在に関する法的な課題等技術以外の問題もあるため,説明可能AIを開発して技術的に解決するべき問題なのかは社会の要請に依存していくものと思われる.

本研究の一部は中部科学技術センター人工知能助成金の助成によるものです.

参考文献

1) 統合イノベーション戦略推進会議決定, 人間中心の AI 社会原則, 総務省ホームページ
2) 内閣府,AI戦略2022,https://www8.cao.go.jp/cstp/ai/index.html
3) M.T. Ribeiro et. al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August, 13-17, 2016, 1135-1144
4) Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.
5) 日本規格協会JIS Z 8115:2019 ディペンダビリティ(総合信頼性)用語
6) 佐藤万企夫, 統計的な考え方からわかること, 鋳造工学第88巻(2016) 第8号 p438
7) Simon, H. A., "Administrative Behavior", 1947.
8) 緒方裕光, リスク解析における不確実性, 日本リスク研究学会, 誌 19(2):3－9(2009)
9) 吉田 満梨, 新市場創造プロセスにおける不確実性と意思決定, マーケティングジャーナル Vol.37 No.4(2018)
10) Koh, Pang Wei, et al. "Concept bottleneck models." International conference on machine learning. PMLR, 2020.
11) Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems 30 (2017).
12) J. van Amersfoort et. al, “Uncertainty Estimation Using a Single Deep Deterministic Neural Network”, ICML 2020.
13) P. Khosla et. al., Supervised Contrastive Learning, 34th Conference on Neural Information Processing Systems (NeurIPS 2020)
14) Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. PMLR, 2016.
15) 仲村佳悟, 人工知能は交通ルールを学べるか?~教師なしコンセプト学習と説明性についての考察~, 情報処理学会研究報告, Vol.2024-ICS-213,No.7,1-5,2024
16) Kahneman, Daniel, and Amos Tversky (1979) "Prospect Theory: An Analysis of Decision under Risk", Econometrica, XLVII (1979), 263-291.
17) Dempster, A. P. (1967). "Upper and lower probabilities induced by a multivalued mapping". The Annals of Mathematical Statistics. 38 (2): 325–339.
18) Shafer, Glenn; A Mathematical Theory of Evidence, Princeton University Press, 1976
19) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)

筆　者

仲村 佳悟
DS部AIセンシング2室インカーG
車内外のAIセンシング技術の開発に従事