特集
Special Feature

［AI 判断の根拠を説明する XAI を使いこなす］

6 信頼できる AI の実現に向けて
― XAI による根拠の納得感向上のアプローチ―

恵木正史　間瀬正啓　濱本真生
（株）日立製作所 研究開発グループ

Shapley 値による根拠説明

機械学習モデル（AI）の予測根拠を説明する方式として，説明の観点，モデルの種別，データの種別によってさまざまな手法が提案されている．本稿では，その中でもデファクトの 1 つとして考えられている Shapley 値による根拠説明方法にフォーカスし，その使いこなしの過程で遭遇した課題とその解決に向けた我々の取り組みについて述べる．

Shapley 値の利用イメージ

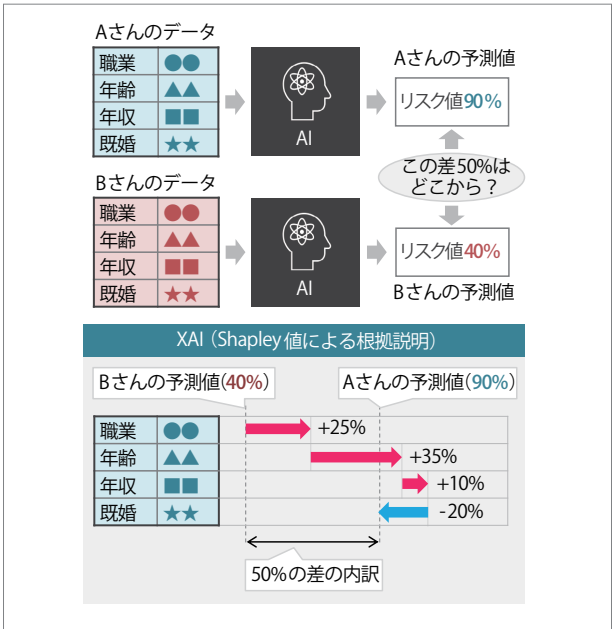
Shapley 値による根拠説明方法は，主に回帰・識別問題の予測モデルを対象に，予測結果がなぜそうなったのかを，入力データの各特徴量がどれだけ貢献したか，といった観点から説明する手法である（図 -1）．

たとえば住宅ローンの与信審査モデルが，申込者 A さんの申請内容（職業＝○○，年齢＝△△，……）に対して，A さんが返済難になる確率が 90% と予測したとする．一方，標準的な申込者 B さんの申込内容（職業＝●●，年齢＝▲▲，……）に対して，予測モデルは B さんが返済難になる確率が 40% と予測したとする．このような場面で Shapley 法は A さんと B さんの予測値の差 50% がいったいどこから来たのかを A さんと B さんの申請項目の差異に基づいて合理的に説明する．

たとえば「A さんの職業が○○であることが確率を 25% 上げる方向に働き，年齢が△△であることが確率を 35% 上げる方向に働き，……，それらの合計値が B さんとの差分の 50% になる」といった貢献度の和として説明する．このような貢献度を Shapley 値と呼ぶ．

なぜ Shapley 値なのか

なぜ Shapley 値による根拠説明方法が支持を集



■図 -1　Shapley 値の利用イメージ

めているかというと，その理由の 1 つは根拠としての数学的な筋の良さだと考えられる．

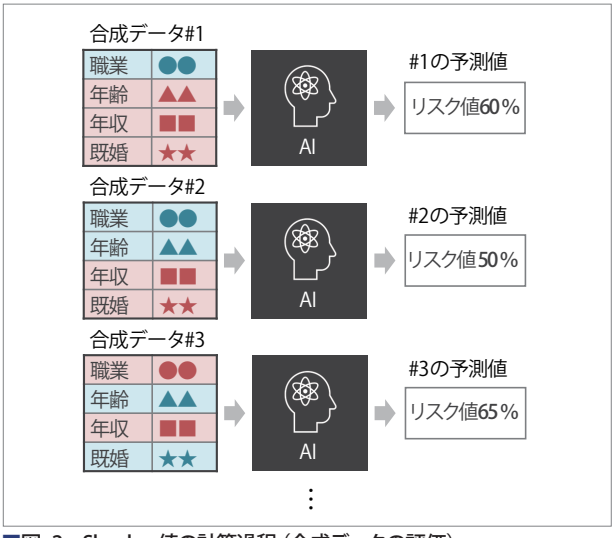
当たり前であるが，モデルの予測値は明確な概念であり，正解値との比較による客観的な評価も容易である．これに対し，根拠は曖昧な概念であり，客観的な評価が難しい．そのため，どうとでも定義できてしまう問題点がある．実際，XAI の分野では一見すると根拠のようではあるが，なぜその方式で計算した値が根拠と言えるのか，が脆弱な研究も多数見受けられる．

この問題に 1 つの答えを出したのが Štrumbelj らによる，Shapley 値を活用した根拠説明方法である [1]．Štrumbelj らは，誰もが合意するであろう「理想的な根拠」が満たすべき 4 つの要件を挙げ，その要件を満たす唯一の解が，古くから協調ゲーム理論で知られた Shapley 値であることを証明した．

すなわち，この方法で得られた値は「理想的な根拠」としての要件をすべて満たすことが数学的に保証されているのである．

Shapley 値の計算方法の概略

以降の解説のために，どのように Shapley 値を算出するのかの概略を説明する．



■図 -2　Shapley 値の計算過程（合成データの評価）

Shapley 値を得るためには，A さんと B さんのデータからさまざまな合成データを作成し，それらに対するモデルの予測値を取得する必要がある（図 -2）．たとえば，B さんの入力データのうち職業だけが A さんの値に置き換わった場合（#1）や，B さんの入力データのうち職業と年齢が A さんの値に置き換わった場合（#2），などである．したがって，入力データの特徴量が N 個ならば，合成データは 2^N 通りある．

これらの合成データとモデルによる予測値の組を，Shapley 値の公式に代入することで，A さんと B さんの予測値の差分 50% を各特徴量の貢献度の和に分解できる．

汎用性の高さ

ここまでの議論から想像されるように，この方法は予測モデルの入力と出力しか使わないため，モデルの内部構造や実装には依存しない．そのため，予測型の機械学習モデルであれば，汎用的に適用することができるというメリットがある．一方で 2^N 通りの計算が必要であるため，計算量が大きいというデメリットがある．この問題に対し，Lundberg らが開発した SHAP 等の OSS は，モンテカルロ法やモデルの内部構造を利用した高速な近似解法を提供している．

Shapley 値活用の勘所

本章ではもう一歩踏み込んで，Shapley 値の活用方法について述べる．

標準的なデータの選び方

前章では A さんの予測根拠を得るために，標準的な申込者 B さん 1 人を基準とした．しかし，実際のユースケースでは，標準的な申込者という 1 人のデータを選定するのは難しい場合も多い．

そのような場合には，標準的な集団を選定して

もよい（図 -3）．この場合 Shapley 値の計算方法は，集団の各人を基準として A さんの Shapley 値を計算し，最後にそれらの平均値を算出すればよい．

重要なのは，A さんの予測根拠を知りたいときに，いったいどのような集団と比較するのかを明確にすることである．たとえば，医療分野で患者の死亡リスクを予測するモデルにおいて，医師が患者 C さんの重症化リスクの根拠を知りたい場合，医師としてどのような集団との比較を求めているか，である．C さんと同じ病気を罹患している人の集団なのか，あるいは C さんと同じ病気を罹患していて，かつ同年代の集団なのか．これは現場のユースケースに依存するため，慎重に選ぶ必要がある．

根拠特性の理解

ここまでは，A さん 1 人の予測結果に対する根拠説明について述べた．このような説明は，個々の予測結果に対する説明であるため，局所説明，アウトカム説明，インスタンス説明とも呼ばれる．

この局所説明を 100 人分，1,000 人分と行っていくと，予測モデルが，どのような場面で何を重視するかといった判断の傾向，すなわち根拠特性が見えてくる．これらは局所説明と区別して，大域説明と呼ばれる．たとえば，年齢を横軸に，縦軸に年齢の Shapley 値をプロットすれば，予測モデルが年齢と予測値の関係をどのように理解し，予測に役立てているかが分かる（図 -4）．こうした大域説明は，学習時に予測モデルが訓練データから発見した知識と解釈できる．

大域説明により，従来は難しかった 2 つのことが可能になる．

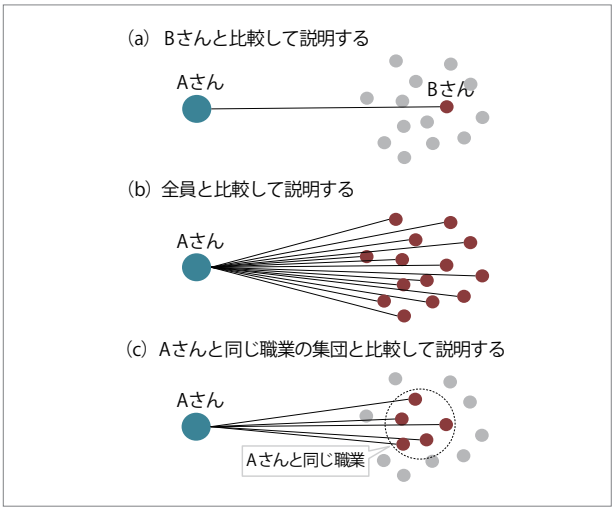
1 つ目は，モデルの知識と専門家の知識とが整合しているかを確認できることである．常に専門家の知識が正しいとは言いきれないが，専門家が「ここは譲れない」という部分で食い違うと，予測モデルに対する信頼を得ることは難しい．

2 つ目は，大域説明を利用することで，モデルの挙動を人が大まかに予測できることである．与信審査の例であれば，A さんが具体的に何をすれば，返済難となるリスク予測値を目標値まで下げることができるのか，その手がかりを得ることができる．

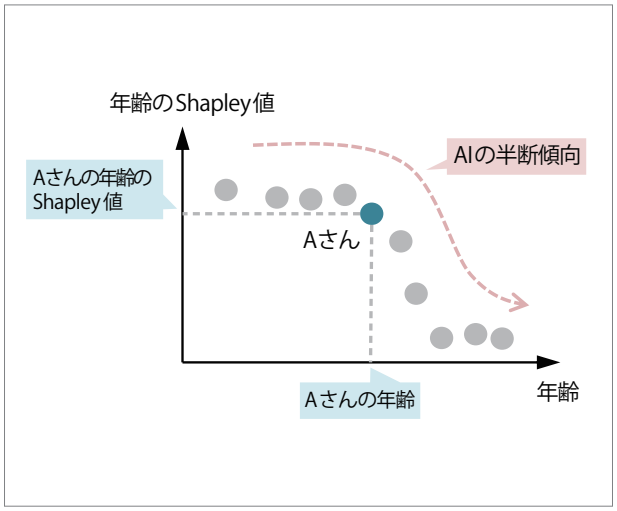
特に 1 つ目は，予測モデルを利用する現場の方々から信頼してもらうためには，不可欠の機能である．

遭遇した課題と解決方法

本章では Shapley 値の使いこなしにおいて直面した 2 つの重要な課題について述べ，我々の解決



■図 -3　標準的なデータの選び方



■図 -4　大域説明

方法について述べる．

課題 1：現実世界にはないデータの発生

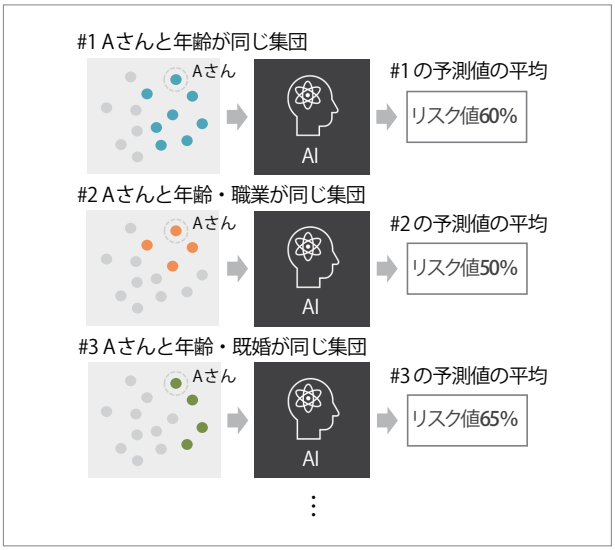
何が問題か

先に述べた与信審査モデルを例に，A さんの Shapley 値の計算のために，標準データ B さんとの合成データを作成し，モデルに予測させる工程について述べた．

ここで 1 つの問題に直面する．合成データは機械的に生成されるので，現実世界の相関を無視したデータや不可能なデータが発生してしまうのである．

たとえば，高給の職業・職位なのに年収が低い，運転免許がないのに職業が運転手，年齢より大きな勤続年数などである．このような異常なデータについて，モデルが予測値を算出したとしても，何か意味を持つとは思えない．逆にそのような不自然な予測値を根拠の計算に組み込むことに対して不信感を生む．さらに，モデルによっては入力項目の組合せをチェックしてエラーを返す場合もあり，予測値が得られないケースもある．

従来の Shapley 法はこのような状況の発生を想定していない．すなわち，現実世界の入力データでは，項目間にはさまざまな相関や制約が入っているのだから，それらを無視して計算された Shapley 値は信頼できるのだろうか？という問題提起である．



■図 -5　Cohort Shapley の説明図

解決策― Cohort Shapley

そこで我々は Stanford 大学の統計学科 Art B. Owen 教授らとの共同研究で，この問題を克服する手法を開発した（図 -5）．提案手法では，合成データを作らずに Shapley 値を出すことができるため，先ほどの問題を回避できる．

「唯一の解」という証明はどうなったのか，という声が聞こえてきそうだが，実は Štrumbelj らの方法には暗黙に仮定されている部分があり，そこには自由度が残っていた．我々はその自由度を活用して，「理想的な根拠」としての要件を満たしながら，合成データを作らなくても Shapley 値を算出できる手法を開発した．その詳細は本稿のスコープから外れるので詳細は割愛するが，誤解を恐れずに言えば，合成データを生成するという操作を，標準的な集団から検索する操作で代替するアプローチである．すなわち，検索されたデータによるモデルの予測値の平均を用いて，Shapley 値を計算する．

現実世界にあり得ないようなデータは登場しないため，入力データの項目間に潜在する暗黙の相関や制約をありのまま受け入れ，その上で Shapley 値を計算することができる．

我々はこの方法を，集団（Cohort）から計算するという意味を込めて Cohort Shapley と呼んでいる [2]．ツールを OSS として公開しているので，興味のある方は使ってみていただきたい☆1．

課題 2：モデルと専門家の知識の不整合

何が問題か

先に述べた通り，モデルが訓練データから学習した知識と専門家の知識が整合しないことがある．たとえばコンクリート強度をセメント量や水分量等から予測する問題において，専門家は，コンクリー

☆1　https://github.com/cohortshapley/cohortshapley

ト強度がセメント量に正比例することや，影響力のある特徴量を経験的に知っていたとする．このとき，予測モデルの根拠特性が正比例とは大きく異なる曲線になったり，専門家が重要と考えていた特徴量の貢献度が上位になかったりすると，専門家はモデルの信頼性に対して疑問を呈するだろう．

つまりモデルと専門家の知識が整合しないと，現場からモデルを信頼してもらえないという深刻な問題が発生する．実際，我々はこのような場面に何度も遭遇した．

この問題をもう少し厄介にしている現象がある．学習条件を変えて学習し直すと，予測モデルの根拠特性が変化するという現象である．これは，同程度の精度を持ちながら，予測結果や根拠特性が異なるモデルが幾通りも存在し得ることを意味する．この現象は，黒澤明監督の映画『羅生門』になぞらえて「羅生門効果：Rashomon effect」と呼ばれている．映画の中では，殺人事件に対して複数の登場人物が異なる証言をして，捜査が行き詰まるエピソードがあるが，我々も同様にどのモデルが一番良いのか，を絞り込めないことを意味する．このような現象は，訓練データが十分でない場合に特に顕著となる傾向がある．

解決策―根拠校正

この問題に対して，我々は羅生門効果を逆手にとれば，専門家の知識との不整合を解決できることに気が付いた[3]．

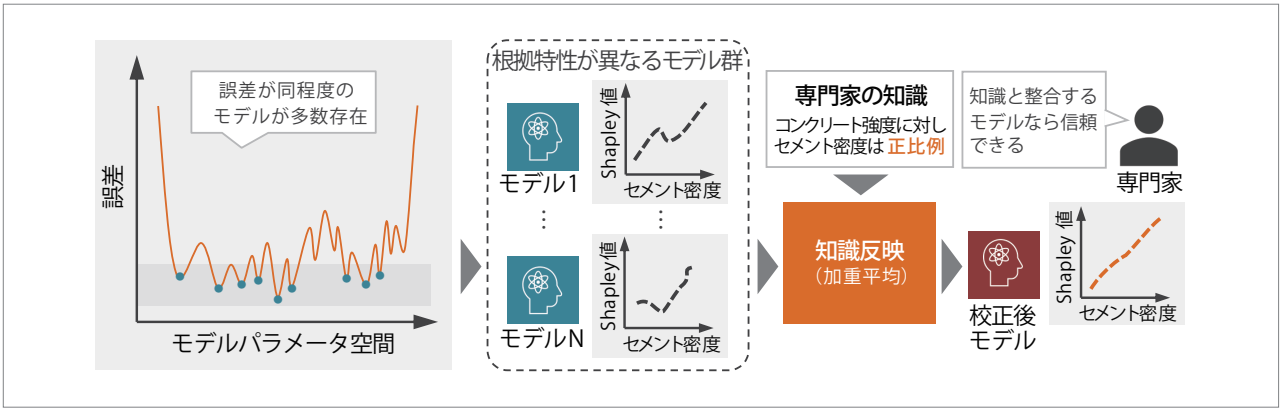
つまり，根拠特性が異なるモデルを多数生成し，それらを適切な重みで線形結合したモデルを作ることで，専門家が考える根拠特性に近づけようというアプローチである．我々はこの方法を根拠校正技術と呼んでいる（図-6）．

ここで強調しておきたいのは，モデルの知識に対して校正を掛けるのは専門家が指摘した部分だけ，という点である．モデルは学習時に訓練データからいろいろな知識を発見するが，そこには専門家もよく知っている知識もあれば，新たな発見もあるだろう．その中で専門家から見ておかしいという部分があれば，そこだけを校正しようという方針である．言い方を変えると，この技術は専門家の知識をモデルに反映する方法の 1 つであると言える．

この技術は，専門家の納得感を高めることにも貢献するが，同時に専門家の知識をモデルに反映することで，予測精度の向上も期待できる．実際，先程のコンクリートの例では，セメント量の根拠特性が正比例に修正されたことで，精度が改善する効果が見られた[3]．

現場からの信頼獲得に向けて

現場の専門家から信頼される AI を実現するには，AI の根拠が現場の知識と整合していることがとて



■図-6　根拠校正の説明図

も重要であると考えている．しかし現場の知識を棚卸すること自体が難しいタスクである．このような場面で，XAI はデータから発見した知識の候補を列挙してくれるので，それを呼び水に専門家からさまざまな知識を聞き出すことが可能となる．その過程で浮き彫りになったギャップを 1 つ 1 つ改善していくことで，真に現場から信頼される AI を実現できると考えている．

今後はそのようなプロセスを効率化する技術の発展が求められると考えられる．本稿がその一助になれば幸いである．

参考文献

1）Štrumbelj, E. and Kononenko, I. : An Efficient Explanation of Individual Classifications Using Game Theory, The Journal of Machine Learning Research, Vol.11, pp.1-18 (2010).
2）Mase, M., Owen, A. B. and Seiler, B. B. : Explaining Black Box Decisions By Shapley Cohort Refinement, Technical Report, arXiv:1911.00467 (2019).
3）Hamamoto, M. and Egi, M. : Model-agnostic Ensemble-based Explanation Correction Leveraging Rashomon Effect, 2021 IEEE Symposium Series on Computational Intelligence, pp.1-8 (2021).

（2022 年 4 月 28 日受付）

■恵木正史　masashi.egi.zj@hitachi.com

（株）日立製作所 研究開発グループ 主管研究員．機械学習モデルの説明性・信頼性・公平性にかかわる研究に従事．

■間瀬正啓（正会員）　masayoshi.mase.mh@hitachi.com

（株）日立製作所 研究開発グループ 主任研究員．機械学習モデルの説明性・信頼性・公平性にかかわる研究に従事．

■濱本真生　masaki.hamamoto.qg@hitachi.com

（株）日立製作所 研究開発グループ 主任研究員．機械学習モデルの説明性・信頼性・公平性にかかわる研究に従事．