最適数値相関ルールを利用したSHAPの予測モデル解釈の補完

山下　史紘†　　亀井　清華†　　森本　康彦†

† 広島大学情報科学部　〒 739–8511 広島県東広島市鏡山 1 丁目 3-2
E-mail: †{b193313,s10kamei,morimo}@hiroshima-u.ac.jp

あらまし　近年, 高精度の予測を実現するニューラルネットワークやアンサンブル学習等のブラックボックスモデルが多くの場面で利用されている. しかしモデルの説明性を伴わず, その不透明性からブラックボックスモデルと呼ばれている. そのようなブラックボックスモデルの説明性を高める技術として eXplainable AI (説明可能な AI) の発展が進んでいるが, その説明はあくまでもモデル解釈の範囲に留まり, 説明と結果の間に因果関係はない. その点から XAI の説明は不十分であると考え, 説明と結果の間にある因果関係を明らかにすることで XAI の説明性を補完する手法を本論文で提案する. 本研究では数値属性のデータを対象とし, 数値属性に対する代表的な XAI である SHAP の説明を補完する. SHAP は各特徴量の予測への貢献度を示す技術であり, インスタンス毎に予測の説明を行うことが出来る. 本研究ではデータマイニング手法の一つである最適数値相関ルールを利用して, SHAP の説明に不足している事実関係の観点を補完した.

キーワード　XAI, SHAP, 相関ルールマイニング

1 はじめに

近年, AI モデルは高精度の予測が可能になり社会的に広く使用されるようになっている. ビジネスの場において使われることも非常に多く, 普段触れているシステムの中にも AI モデルを利用しているシステムは多く存在する. そのような場面において AI モデルが一番に求められることは予測精度である. しかし, 社会的に重責を担うようなシステムに AI モデルが組み込まれる場合, 単に精度が高いだけでは実用性に欠ける. 例えば, AI による自動採点システムは正確であることは前提として, 不正解を付けた個所に対しては説明を求められる場合があるだろう. このような「説明」に関する諸問題を解決する手段として XAI 技術が注目を集めている.

XAI は様々な形式のデータに対応しているが, 本論文では数値属性のテーブルデータを扱う. 代表的な XAI 技術として SHAP [1] が挙げられる. SHAP による説明からは各特徴量の予測への貢献度を知ることができる. しかしこの説明はあくまでモデルに対して行われているものであり, 説明を受け取る場合, その点に注意しなければならない. (例 : ローン審査の AI モデルに対して XAI 技術を使用した結果, 審査不合格の予測に大きく影響したのは「“年収” = 200 万円」であった. この時, 年収 200 万円の人はローン審査に合格しない, というように事実関係として解釈するのは誤りであり, あくまでモデルがその特徴量を重視したというモデルの解釈に留める必要がある.) この点から, XAI による説明は不十分であると考える. そこで事実関係として解釈できる相関ルールを利用して, XAI による説明を補完する手法を提案する.

2 関連研究

2.1 相関ルール

$\mathcal{I} = \{I_1, I_2, ..., I_m\}$ を 2 値属性のアイテム集合とする. $T = \{t[1], t[2], ..., t[n]\}$ をトランザクションのデータベースとする. 各トランザクション t は 2 値ベクトルとして表現され, t がアイテム I_k を購入した場合は $t[k] = 1$, そうでない場合は $t[k] = 0$ とする. データベースには各トランザクションについて 1 つのタプルが存在する. X は $\mathcal{I}$ の部分集合であり, X の中のすべてのアイテム I_k に対して $t[k] = 1$ のとき, トランザクション t は X を満たすと言う. [5] ただし, $I_j \notin X$. このとき, X を満たすトランザクション t が I_j も満たすという規則を相関ルールといい, 次のように記述する [5].

$$X \Rightarrow I_j$$

相関ルールの有意性を示す指標として支持度と確信度があり, それぞれ次のように定義されている.

支持度 : ルールの適用性を表し, T の内 X, I_j を共に含む確率.
確信度 : ルールの強さを表し, T の内 X を含む全トランザクションにおいて X, I_j を共に含む条件付確率.

相関ルールを抽出する相関ルールマイニングにおいては, ユーザは任意に最小支持度と最小確信度を設定し, それらを共に満たすルールを抽出する. ただし, 相関ルールはカテゴリカルデータにのみ対応しており数値データを扱うことは出来ない.

2.2 最適数値相関ルール

現存するデータベースの多くは数値属性の特徴量を含んでおり, 数値属性に対応した相関ルールマイニング手法 [6] を本節で紹介する.

A をデータベースにおける数値属性の特徴量とし, I をある数値区間とする. 2 値属性の特徴量 Y に対して, 特徴量 A の値

が I の範囲にあるとき Y が起こるという規則を最適数値相関ルールといい, 次のように記述する.

$$(A \in I) \Rightarrow Y$$

また, データベースのタプル数を N, $A \in I$ を満たすタプル数を s, ルールに該当するタプル数を h とする. これらを用いて最適数値相関ルールにおける支持度と確信度は次のように定義できる.

支持度： $\frac{h}{N}$　　確信度： $\frac{h}{s}$

最適数値相関ルールには支持度最適化ルールと確信度最適化ルールがあり, 本研究では SHAP の説明の補完に確信度最適化ルールを利用する.

最適化数値相関ルールの “最適化” とは, 区間 I を最適化することを意味する. 確信度最適化ルールは最小支持度を任意に設定し, それを満たしたうえで確信度を最大化する区間 I_{max} を持つルールを求める問題に等しい. アルゴリズムを簡単に説明する.

与えられたデータベース R のタプルを t とし, そのタプルの数値属性 A の値を $t[A]$ と表記する. 属性 A の定義域を次のような交わりのないバケットの列 $B_1, B_2, ..., B_M (B_i = [x_i, y_i], x_i \leqq y_i < x_{i+1})$ に分割し, 全てのタプルの属性 A の値が必ずどれかのバケットに入るように分割する. 集合 $\{t \in R | t[A] \in B_i\}$ に入るタプル数を B_i のサイズといい, u_i とする. また, 集合 $\{t \in R | t[A] \in B_i, t[Y] = 1\}$ に入るタプル数を v_i とする. このとき, 支持度は $(\sum_{i=p}^{q} u_i)/(\sum_{i=1}^{n} u_i) \sim \sum_{i=p}^{q} u_i$, 確信度は $(\sum_{i=p}^{q} v_i)/(\sum_{i=p}^{q} u_i)$ と表現できる. バケットの列のイメージを図 1 に示す. 支持度が最小支持度以上になるような任意の区間において, 最も確信度が高くなる区間が確信度最適化ルールとなる.

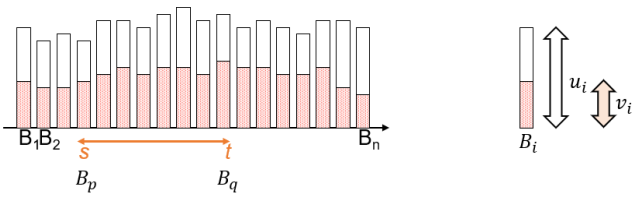

図 1　バケットの列のイメージ

2.3　SHAP

SHAP [1] はゲーム理論において各プレイヤーの寄与度を算出するために利用される Shapley 値を基にした XAI 技術であり, 予測に対する各特徴量の寄与度をインスタンスごとに算出する [7]. 図 2 は後述のデータセットにおいて, 糖尿病であると正しく予測したインスタンスに対する SHAP による説明である. 赤枠は正方向の影響の大きさを表し, 青枠は負方向の影響の大きさを表す. 赤色の特徴量は正の, 青色の特徴量は負の Shapley 値を持ち, 各特徴量の Shapley 値の合計が予測値 f(x) となる. このインスタンスに与えられた説明は, “Glucose = 184” であることが糖尿病と予測した一番の根拠であると解釈する. このように, SHAP はインスタンス毎に説明を与えるミクロな XAI 技術である.

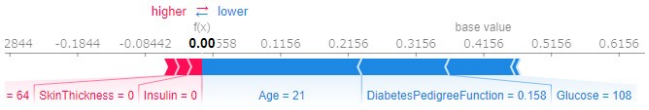

図 2　SHAP による説明

3　提 案 手 法

本節では SHAP による説明を最適数値相関ルールを用いて補完する手法について述べる.

はじめに, データセットを訓練データとテストデータに分割し予測モデルを作成する. この際, 採用するモデルは任意であるが正解率が高くなるようにモデルを選択する. 次に, テストデータの中から正しく予測できていたインスタンスを抽出し, それらに対して説明を生成する. 説明の表示には SHAP の force plot を利用する.

生成した説明から最も Shapley 値が高い特徴量をインスタンス毎に抽出し F_{top} とし, F_{top} とそれが取る値から成るキーバリューペアを作成する. 図 2 のインスタンスにおいては $F_{top} =$ Age であり, キーバリューペアは [Age： 21] と記述する. 次に, インスタンス毎に F_{top} に関する最も確信度の高いルールを抽出する. 図 2 のインスタンスにおいては, Age に関する 21 を含む区間 I を持つルールの内, 最も確信度の高いルールは以下の通りであった.

$$(Age \in [21, 22]) \Rightarrow (Outcome = 0)$$

確信度： 0.847

最小支持度： 20

以降では最も確信度の高いルールを R_{top} とする.

抽出したルールの区間と確信度を用いて SHAP による説明を補完する. 例として図 2 のインスタンスの場合, 「Age の値が 21 から 22 の人の内 84.7%の人が糖尿病ではない」というデータセットにおける事実関係を提示する. このように, SHAP の説明に事実関係の尺度を導入することで, SHAP による説明と予測結果の因果関係を解釈することが出来る.

4　実　　験

本節では提案手法による実験結果を報告し考察を述べる.

本研究では Kaggle[1]にて公開されている, 糖尿病予測 [2], 乳がん予測 [3], の 2 つのデータセットにて実験を行った. また, 機械学習モデルはどちらのデータセットに対しても LightGBM を用いた.

1：https://www.kaggle.com/

2：https://www.kaggle.com/datasets/whenamancodes/predict-diabities

3：https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

図 3　糖尿病予測インスタンス 1

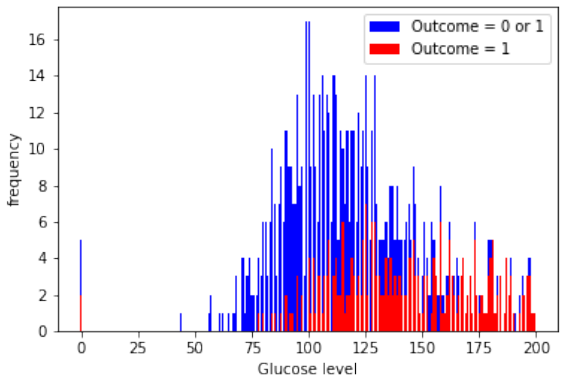

図 4　糖尿病患者数の分布 (Glucose)

4.1　結果と考察

各データセットに対する予測モデルの正解率は, 糖尿病予測モデルが 72.08%, 乳がん予測モデルが 96.49%であった. 本論文ではデータセットごとに, いくつかのインスタンスに対する結果を示す.

4.1.1　糖尿病予測

糖尿病予測データセットは Glucose(血液中のグルコース濃度), Age(年齢), BMI(肥満度) 等の 8 個の説明変数を用いて, Outcome(0 ならば非糖尿病, 1 ならば糖尿病) を予測するものである.

図 3 のインスタンスは F_{top} = “Glucose” である. また, 特徴量 “Glucose” のバケット列を図 4 に示す. これは糖尿病患者数の分布を表しており, 糖尿病患者数を赤, 糖尿病患者数と非糖尿病患者数の合計を青で表している. [Glucose : 184] に関する R_{top} と確信度を以下に示す.

$$R_{top} : (Glucose \in [167, 189]) \Rightarrow (Outocome = 1)$$
確信度 : 0.895
最小支持度 : 20

図 3 の SHAP による説明からはグルコース値が予測に大きく影響したということまでしか読み取ることは出来ないが, R_{top} を補足することによって, 「グルコース値が 167 から 189 の人の内, 89.5%の人が糖尿病である」という事実的な観点が補われた. この R_{top} と図 4 から分かるように, グルコース値が高いほど糖尿病になりやすい. 図 3 から “Glucose” の Shapley 値が突出して高いことから, 予測モデルもこの特徴を予測に反映しているのではないかと推測できる.

次に, 図 5 の説明を補完する. このインスタンスは F_{top} = “Age” である.　また, 特徴量 “Age” のバケット列を図 6 に示す. [Age : 56] に関する R_{top} と確信度を以下に示す.

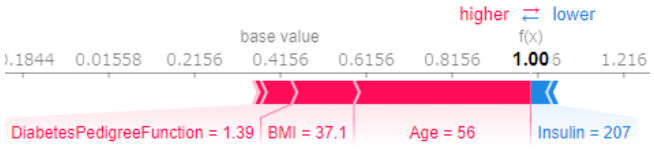

図 5　糖尿病予測インスタンス 2

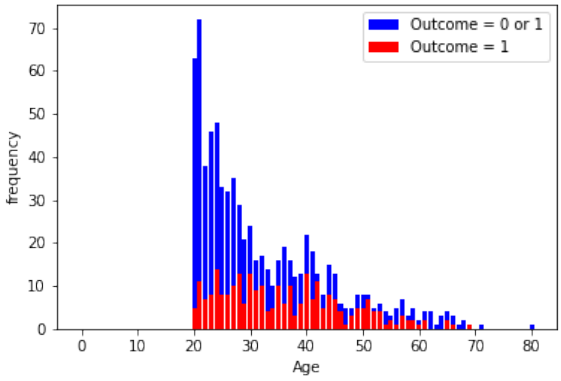

図 6　糖尿病患者数の分布 (Age)

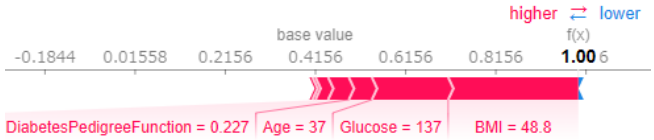

図 7　糖尿病予測インスタンス 3

$$R_{top} : (Age \in [51, 56]) \Rightarrow (Outcome = 1)$$
確信度 : 0.692
最小支持度 : 20

図 5 の SHAP による説明からは Age が予測に大きく影響したということまでしか読み取ることは出来ないが, R_{top} を補足することによって, 「年齢が 51 歳から 56 歳の人の内, 69.2%の人が糖尿病である」という事実的な観点が補われた. SHAP の説明のみでは年齢による糖尿病への影響を考察することは出来なかったが, R_{top} を補足することで 50 代前半の糖尿病率が 70%近くあるという情報が加わり, 年齢と糖尿病の関係についての情報を得ることが出来る.

図 7 のインスタンスは F_{top} = “BMI” である. また, 特徴量 “BMI” のバケット列を図 8 に示す. [BMI : 48.8] に関する R_{top} と確信度を以下に示す.

$$R_{top} : (BMI \in [46, 56]) \Rightarrow (Outocome = 1)$$
確信度 : 0.636
最小支持度 : 20

これは「BMI の値が 46 から 56 の人の内, 63.6%の人が糖尿病である」という事実を補足している. 先に紹介した 2 つのインスタンスに比べると確信度が低い結果となったが, バケットの分布による理由が考えられる. 図 8 より, BMI の値が 48.8 周辺

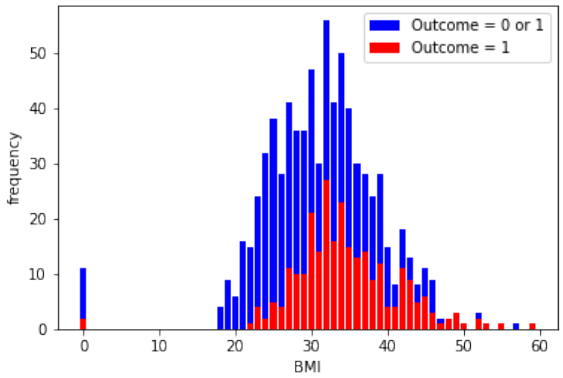

図 8　糖尿病患者数の分布 (BMI)

において糖尿病患者の割合は高い．しかし, 人数が極端に少なくなっている．R_{top} を導出する際に最小支持度は 20 とした．そのため, 最小支持度を満たすために数値区間の下限が下がり, 非糖尿病患者の割合が高いバケットを含んだため確信度が下がったと推測できる．

続いて非糖尿病であると正しく予測したインスタンス (図 9) の説明を補完した結果を報告する．このインスタンスは F_{top} =“Glucose” であり, キーバリューペアは [Glucose： 74] となる．また, 図 10 に特徴量 “Glucose” のバケット列を示しているが, 非糖尿病であるインスタンスのため, 非糖尿病患者数を赤, 糖尿病患者数と非糖尿病患者数の合計を青で表している．得られた R_{top} と確信度は次の通りであった．

$$R_{top} : (Glucose \in [50, 80]) \Rightarrow (Outcome = 0)$$
確信度： 0.971
最小支持度： 20

図 9　糖尿病予測インスタンス 4

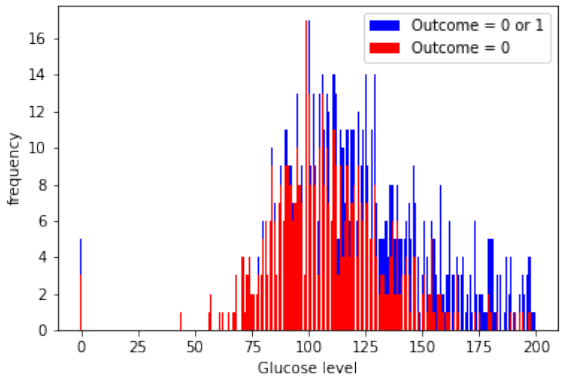

図 10　非糖尿病患者数の分布 (Glucose)

図 11　乳がん予測インスタンス 1

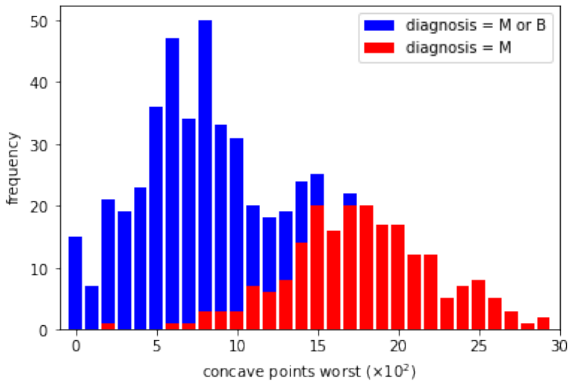

図 12　悪性腫瘍の分布 (concave points worst)

SHAP の説明のみでは, グルコース値が予測に大きく影響したということまでしか解釈できないが, R_{top} を補足することによって,「グルコース値が 50 から 80 の間の人の内, 97.1%の人が糖尿病でない」という事実を知ることが出来る．これにより, グルコース値が低い人のほとんどは糖尿病ではないため, 図 9 のインスタンスは糖尿病ではないと予測されたと, 因果関係として解釈できるようになる．

4.1.2　乳がん予測

続いて, 乳がん予測データセットに対する実験結果を報告する．このデータセットは腫瘍が悪性か良性であるかを予測するものである．1 つの腫瘍を様々な方向から撮影した画像から concave points(輪郭の凹部の数), area(面積), perimeter(外周長) 等, 全 10 項の mean(平均値), worst(最悪値), se(誤差) を計測した 30 の特徴量から成る．目的変数は diagnosis であり, M ならば悪性, B ならば良性 [4]を意味する．

図 11 のインスタンスは “concave points”(輪郭の凹部の数) の最悪値が 0.265 であることに最も影響を受けて乳がんであると正しく予測された．F_{top} は “concave points worst” であり, それに伴う悪性腫瘍の分布を図 12 に示す．横軸は腫瘍の輪郭の凹部の数の最悪値を示し, 縦軸は腫瘍の数を示す．悪性腫瘍 (M) の数を赤, 悪性腫瘍と良性腫瘍 (B) の合計数を青で示している．図 11 の説明に対して得られた R_{top} と確信度を以下に示す．

$$R_{top} : (\text{concave points worst} \in [0.18, 0.29])$$
$$\Rightarrow (diagnosis = M)$$
確信度： 1.0
最小支持度： 30

4：SHAP 適用のため M, B をそれぞれ 1, 0 にエンコードしている．

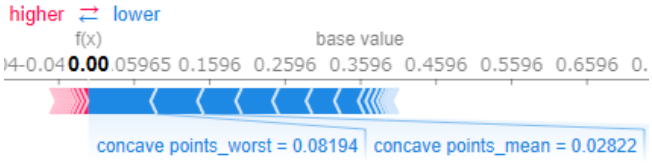

図 13　乳がん予測インスタンス 2

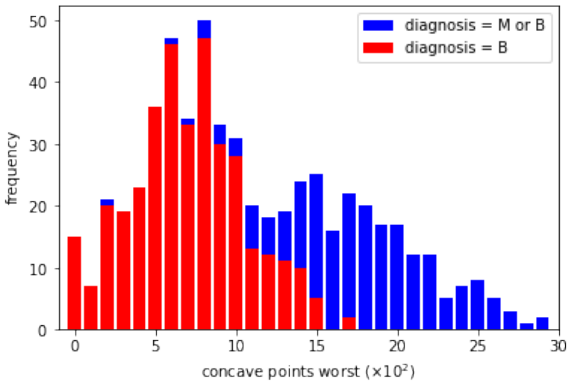

図 14　良性腫瘍の分布 (concave points worst)

これは「輪郭の凹部の数の最悪値が 0.18 から 0.29 の間の腫瘍の内, 100%が悪性腫瘍である」という事実を補足している. 図 12 より, “concave points worst” の値が 0.18 以上の範囲では全て悪性腫瘍である. そのため, 確信度は 1.0 であり, 数値区間 I も 0.18 以上の全インスタンスを含む区間になっている.

次に F_{top} は同じく “concave points worst” であり, 良性腫瘍であると正しく予測したインスタンスを図 13 に示す. また, 良性腫瘍の数を赤で示した良性腫瘍の分布を図 14 に示す. このインスタンスに対して得られた R_{top} と確信度を以下に示す.

$$R_{top} : (\text{concave points worst} \in [0.00, 0.09])$$
$$\Rightarrow (diagnosis = B)$$
確信度：0.976
最小支持度：30

これは「輪郭の凹部の数の最悪値が 0.00 から 0.09 の間の腫瘍の内, 97.6%が良性腫瘍である」という事実を補足している.

悪性腫瘍, 良性腫瘍どちらの場合も, 特徴量 “concave points worst(輪郭の凹部の数の最悪値)” に関するルールは確信度が非常に高かった. これは図 12, 14 から分かるように, “concave points worst” の値が高い範囲では悪性腫瘍, 低い範囲では良性腫瘍の割合が極端に高いからであろう. また, F_{top} =“concave points worst” であったインスタンスは全テストデータ 114 の内, 67 であり, 全特徴量の中で最も多かった. これは機械学習モデルも予測の際に “concave points worst” の値を多くの場合で重視しているということである. モデルもこのような特徴量の特性を学習したのかもしれない.

2 つのデータセットの結果を比較すると, 乳がん予測データセットの方が予測精度も確信度も高かった. 恐らくこれはデータセットの特徴の問題だと考えられる. 図 15, 16 に乳がん予測データセットにおける特徴量 “perimeter worst(外周長の最悪値)”, “area worst(面積の最悪値)” に伴う悪性腫瘍の分布を示す. これらも “concave points worst” に伴う悪性腫瘍の分布 (図 12) と同様の特徴がある. そのため同様の理由から確信度は高くなり, 予測においてはモデルが学習しやすいのかもしれない.

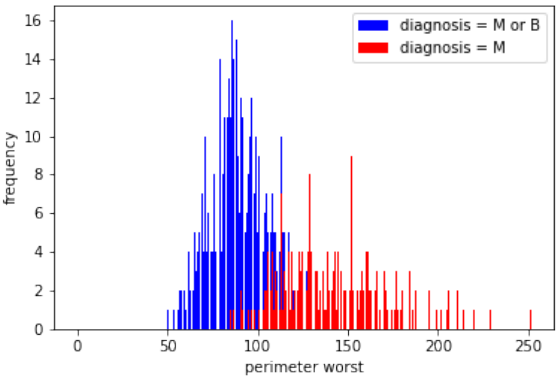

図 15　悪性腫瘍の分布 (perimeter worst)

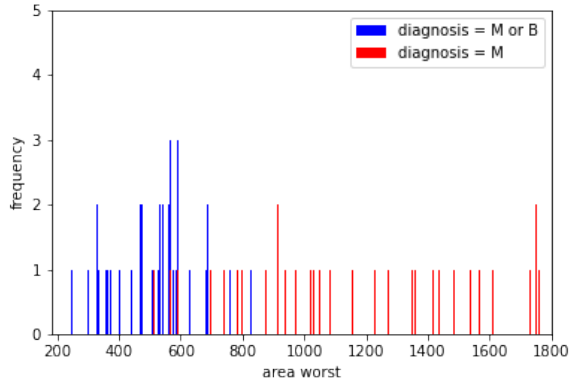

図 16　悪性腫瘍の分布 (area worst)

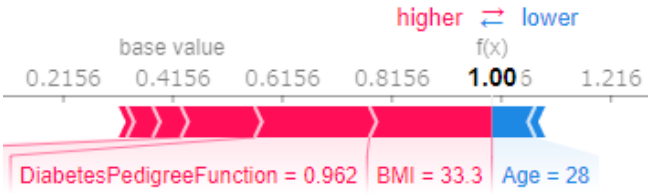

図 17　僅差で F_{top} とならない例 (area worst)

5　おわりに

最適数値相関ルールを用いて SHAP による説明の補完を行った. 糖尿病予測と乳がん予測の 2 つのデータセットに対して実験を行い, 各インスタンスに対して最適数値相関ルールと確信度を導出し, 説明を補完することが出来た. SHAP の説明だけではモデルの解釈に留まるが, 提案手法によって SHAP の説明に不足していた事実関係の観点を補うことが出来た. このような観点の補完により, SHAP による説明の情報量が増加し, よ

り有用な説明になると考えている．

今後の課題としては以下のようなことが挙げられる．まずは特徴量の抽出である．提案手法では Shapley 値が最も大きい特徴量を F_{top} として抽出した．しかしこの抽出方法では僅差で F_{top} とならなかった特徴量の影響を無視してしまう．図 17 のような場合においては, あまり Shapley 値に差がないが特徴量 “Diabetes Pedigree Function(遺伝的影響)” を無視することになる．また, 本研究においては XAI の代表的な技術である SHAP の説明に対してのみ実験を行った．SHAP のようにインスタンス毎の説明を行える他の技術として LIME [9] がある．そのような他の XAI 技術に対する実験を行って提案手法の妥当性を検証していきたい．

謝　　辞

本研究は科研費 20K11830 の助成を受けたものです．

文　　献

[1] Scott M. Lundberg, Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (ed.), Advances in Neural NIPS’17: Proceedings of the 31st International Conference on Neural Information Processing Systems (2017).

[2] Leo Breimen. “Bagging predictors.” Machine learning 24, pp.123-–140 (1996).

[3] Leo Breimen：“Statistical Modeling.: The Two Cultures”, Statistical science. 16 (3) pp.199–231 (2001).

[4] Gunning, David：“Explainable artificial intelligence (xai).”, Defense advanced research projects agency (DARPA), nd Web 2.2 (2017).

[5] Rakesh Agrawal, Tomasz Imielinski, Arun Swami “Mining Association Rules between Sets of Items in Large Databases” Proceeding of the 1993 ACM SIGMOD international conference on Management of data, pp.207–216 (1993).

[6] Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, Takeshi Tokuyama “Mining Optimized Association Rules for Numeric Attributes” Proceedings of the 15th ACM SIGACTSIGMOD-SIGART Syrup. on Principles of Database Systems (PODS ’96), Montreal, Canada (1996).

[7] Christoph Molnar: “Interpretable Machine Learning A Guide for Making Black Box Models Explainable.” https://hacarus.github.io/interpretable-ml-book-ja/

[8] 斎藤重幸: “わが国の糖尿病のトレンド”, 総説（循環器病予防総説シリーズ 10：記述疫学編 4 , (2018).

[9] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin: “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier” KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1135–1144 (2016)