複数の局所的説明の比較による分類モデル解釈支援のための可視化手法

Visualization for Interpreting Classification Models by Comparing Multiple Local Explanations

澤田 頌子 *1　　豊田 正史 *2
Shoko SAWADA　　Masashi TOYODA

*1東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

*2東京大学 生産技術研究所
Institute of Industrial Science, the University of Tokyo

The reliability of machine learning models are essential when using models to make decisions in the real world. Various approaches have been proposed to make the models more interpretable and the prediction results more trustworthy. Among them, methods that regard the models as black box and provide reasons for predictions are effective especially for users who have little knowledge of machine learning, such as domain experts and end users. In this paper, we propose a visual analysis method to support users in evaluating and improving the reliability of a model by utilizing model-agnostic explanation methods. Specifically, we adopt multiple local explanation techniques which explain individual input data and generate local explanations for a large set of data points and visualize them in a heat map. It reveals features that affect a wide range of the model. We applied our visualization to a diabetes classification model and verified its effectiveness in evaluating the trustworthiness of the model.

1. はじめに

近年機械学習モデルはさまざまな分野で高い性能を発揮しており，実世界での利用も進んでいる．その一方で，特に医療や金融といった意思決定の影響が重大な分野では，モデルやその出力に対する透明性や解釈性が求められている．

モデルの解釈性を高めるために多様なアプローチが提案されている．その中でもモデルの複雑な内部構造を数値化・可視化してモデルの透明化を目指すのではなく，モデルをブラックボックスと見なして，人間が理解できる形でモデルが学習した知識や予測根拠を提示する説明手法は，機械学習に関する深い知識を持たない他分野の専門家やエンドユーザも理解できるため，モデルの解釈性を高め，実世界展開を進める際には非常に有効である．さらに可視化やインタラクティブな分析インタフェースを導入してモデル探索や改善を支援する手法も活発に研究されているが，多数の入力データに対するモデルの判断傾向やエラーの原因を理解するのは容易ではない．また多数提案されている説明手法を適切に選択して利用するためには，それぞれの説明手法の特性を理解している必要もある．

本論文では，モデルに依存しない説明手法のうち，個別の事例に説明を付与する局所的な説明手法を用いてモデルの解釈と改善を支援する可視化分析を提案する．大量の入力データに対して局所的な説明を生成し，それをヒートマップや棒グラフの集合で可視化することで，多様な入力に対するモデルの挙動の分析が可能になる．このとき複数の説明手法を採用して可視化することで，それぞれの手法により生成された説明を比較しながらモデルが学習した知識や予測根拠を理解することで，より効率的なモデル解釈・改善が実現することを目指す．

2. 関連研究

本章ではモデルをブラックボックスとして扱い，モデルの入出力を元に説明生成する手法のうち，単一の入力データに対して予測根拠を提示する局所的な説明手法を紹介する．また可視化やインタラクティブなインタフェースを導入してモデルの性能評価や解釈を目指す手法についても紹介する．

2.1 局所的な説明手法

入力データの特徴量を用いる説明手法としては，LIME[Ribeiro 16]，SHAP[Lundberg]，Anchor[Ribeiro 18] が代表的である．LIME では，学習済みモデル f と説明対象の入力データ x が与えられたら，x の周辺で局所的にモデル f に近似した線形モデル ξ を学習し，ξ の特徴量とその重みを説明として利用する．LIME はモデルの予測が Anchor[Ribeiro 18] では，説明対象の入力データがその予測結果を得るために必要な特徴量の条件を多腕バンディットにより特定し，ルールとして提示している．LIME も Anchor も，説明をより解釈可能なものにするため，事前に特徴量を解釈可能なものに設定する必要がある．その他にも，予測根拠として説明対象の入力データと類似する入力データを提示する Influence[Koh 17] や，予測結果を反転させるために入力データに必要な変更を提示する Counterfactual Explanation[Wachter 17] なども挙げられる．

2.2 モデルの可視化分析手法

機械学習モデルをより説明可能なものにするために，さまざまな可視化分析手法が提案されている．Manifold[Zhang 19] や The what-if tool[Wexler 20] は，大量の入力データや予測結果などを可視化して統計分析することでモデルの性能の評価や比較を可能にした．モデルの解釈・改善のためのアプローチとしては，インタラクティブな可視化を提供してモデルが学習した知識や予測根拠をユーザに探索させる手法 [Krause 17][Ming 19][Cheng 20] が代表的である．Spinner らは，モデルの学習から説明手法による評価・改善までを一連のプロセスとしてシステム化した explAIer[Spinner] を提案した．

本手法では，大量の入力データに対して複数の説明を可視化しユーザに提示することで，ユーザの効率的なモデル解釈・改善を支援することを目指す．

連絡先: {shoko, toyoda}@tkl.iis.u-tokyo.ac.jp

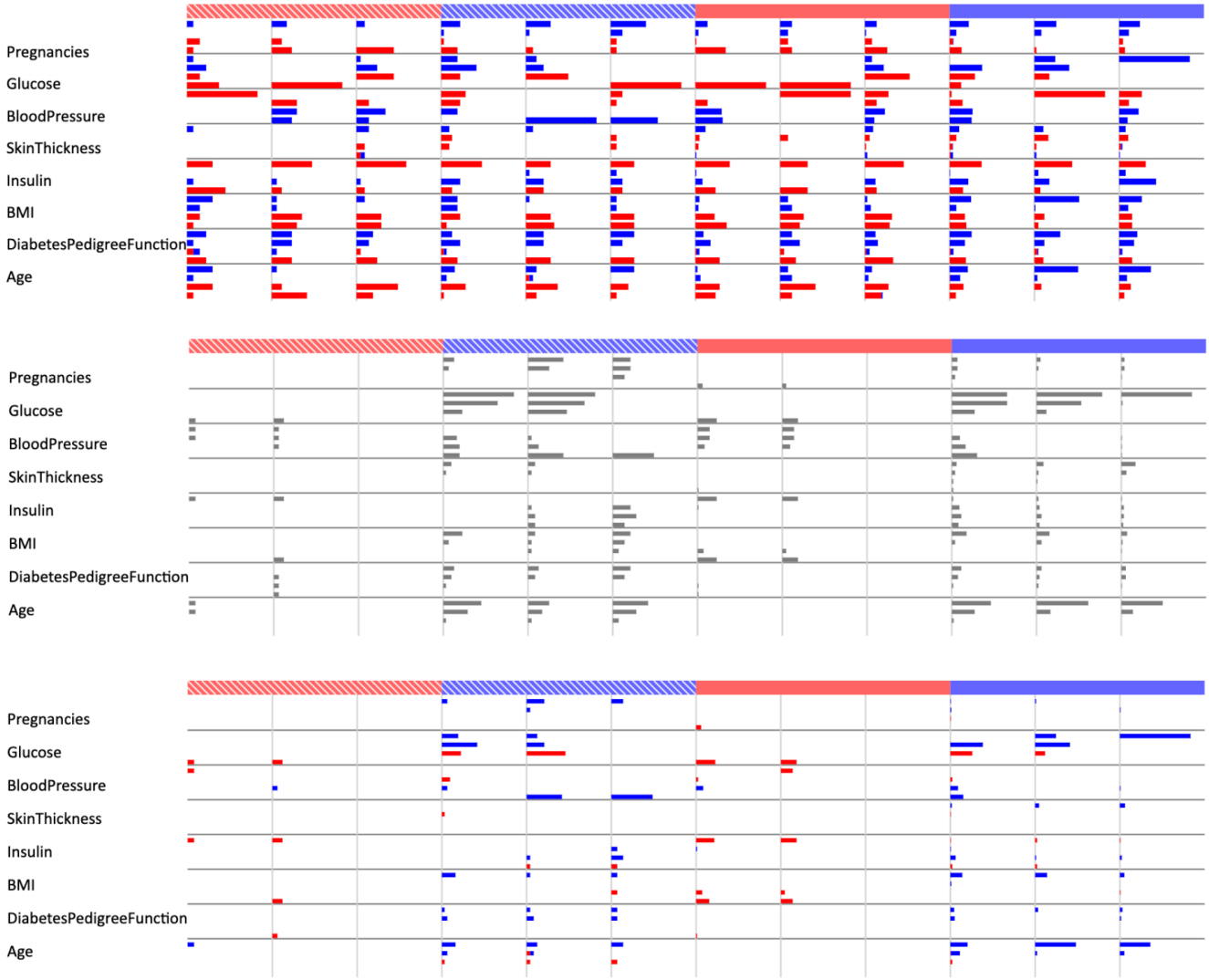

図 1: 要約ビュー (上) LIME の要約 (中) Anchor の要約 (下) LIME と Anchor 両方の要約

3. 提案手法

本章では，複数の説明手法を用いたモデル解釈支援のための可視化について述べる．本手法はモデルが学習した知識や予測に失敗する原因を探索可能な可視化分析を実現するために，大量の入力データに対する説明を要約した要約ビュー，説明の一覧提示するヒートマップビュー，さらに個別の入力データに関するより詳細な情報を確認するためのローカルビューの 3 つの可視化を提供する．図 2 に，本手法の概要を示す．まず全入力データとその予測結果に対して複数の局所的な説明手法により説明を生成し，大規模な局所的な集合を生成する．これを要約ビューやヒートマップビューなどで可視化することで，多様な入力に対するモデルの挙動を分析し，解釈・改善を目指す．

3.1 説明生成

本手法では学習済みモデルと入力データに対して，複数の説明手法を用いてすべての入力データに対して説明を生成する．現在は説明手法として，LIME[Ribeiro 16] と Anchor[Ribeiro 18] の 2 つの特徴量ベースの局所的な説明手法を採用している．2.1 節で述べたように，LIME の説明は特徴量と重みの集合で，Anchor の説明は特徴量で構成されたルールの集合である．このとき，LIME と Anchor により説明が付与された入力データ数を N，これらの説明に登場する特徴量の種類数を F とすると，ここで生成された説明はそれぞれ，$F \times N$ の説明行列として見なすことができる．

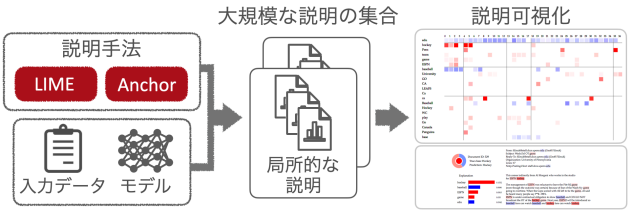

図 2: 提案手法の概要．

3.2 要約可視化

機械学習では入力データ数や特徴量数が膨大になることがあるため，$F \times N$ の説明行列を直接可視化すると煩雑になり，効率的なモデル分析は難しい．そこで本手法では，説明行列の概要として要約ビュー (図 1) を提供する．

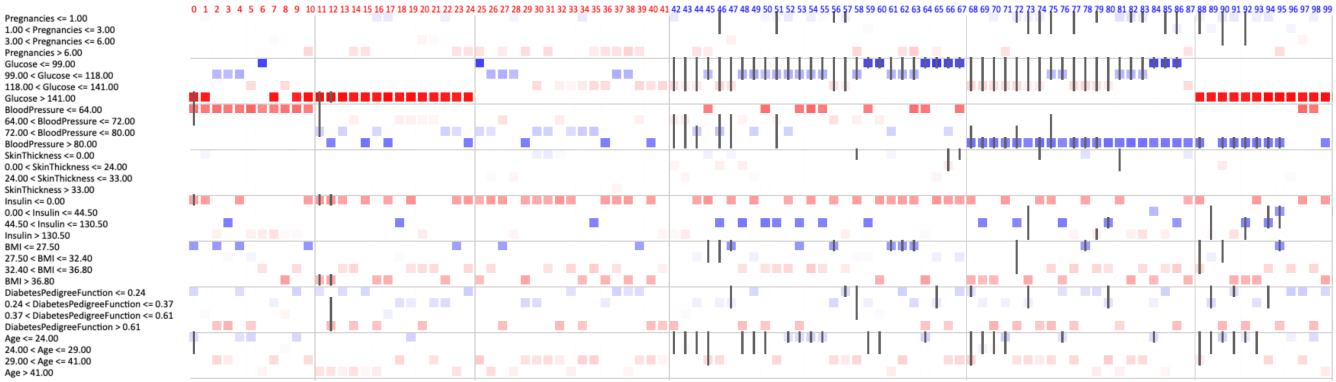

図 3: LIME と Anchor の説明のヒートマップ

要約ビューでは棒グラフが横に多数並んでいるが，これは説明行列をクラスタリングし，各クラスタを棒グラフに要約したものである．モデルが正しく学習した知識や予測ミスの傾向の分析を可能にするため，まず説明行列の入力データを false positive，false negative，true positive，true negative に分け，さらにそれぞれに Spectral clustering[Shi 00] を適用し，類似する説明を持つ入力データを 1 つのクラスタにしている．各棒グラフは，クラスタ内で各特徴量が説明として登場した回数を示していて，行方向に F 個の特徴量が並んでいる．この要約ビューは，表形式データの二値分類タスクに適用した例で，ここでは説明生成に必要な解釈可能な特徴量として，連続値の特徴量を四分位量子化したものを利用しているため，1 つの特徴量に対して 4 つのバーが生成されており，上から特徴量の範囲の値の小さい順で並んでいる．

今回採用した 2 つの説明手法それぞれの説明の傾向を比較することでモデルの効率的な理解につながると考えるため，この要約は説明手法ごとに提供する．つまり，LIME の説明に使われた特徴量の要約 (図 1(上))，Anchor の説明に使われた特徴量の要約 (図 1(中))，そして LIME と Anchor の両方の説明で同時に使われた特徴量の要約 (図 1(下)) の 3 種類の要約が存在する．

現在本手法では二値分類のタスクを対象としているため，2 つのクラス (positive/negative) をそれぞれ赤と青で表現した．LIME の要約の棒グラフでは，positive な重みがついた回数は赤で，negetive な重みがついた回数は青で表示している．棒グラフの上の矩形は色とパターンで混同行列を表していて，色はモデルの予測，斜線はモデルが予測を間違えていることを示している．つまりここでは false positive，false negative，true positive，true negative の順に並んでいる．Anchor の説明における特徴量はすべて予測と同じクラスに寄与していることになるから，棒グラフの色は一貫してグレーで表示する．

3.3 ヒートマップ可視化

要約ビューで興味を持ったクラスタの詳細な探索を可能にするために，クラスタリングした $F \times N$ の説明行列をヒートマップで可視化する．図 3 に，ヒートマップの一部を示す．ここでは false positive のクラスタと false negative のクラスタが並んでいる．行に特徴量が並んでいて，これは要約ビューの特徴量の並びと同一である．各列が各入力データの局所的な説明に対応している．正方形のセルが LIMEn 説明の重みを示していて，要約ビューと同様に赤が positive の予測の根拠，青が negative の予測の根拠で，色の明度で重みの大きさを表している．グレーの矩形は Anchor のルールによりカバーされる特徴量の範囲を示している．クラスタの区切りを垂直方向の線で示している．

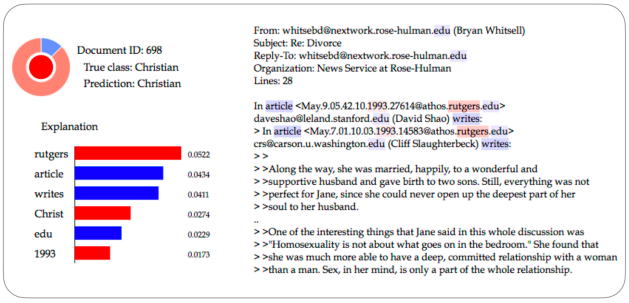

図 4: ローカルビューの一例 (テキスト二値分類タスク)

3.4 ローカルビュー

さらに本手法では，要約やヒートマップによる説明の概要を見て，ユーザが関心を持った個別の入力データについて詳細な情報を提供するためのローカルビューも用意している．ローカルビューは適用するデータやタスクに応じて可視化する内容を変えているが，ここでは一例として図 4 にテキストの二値分類タスクのローカルビューを示す．図 4(左) ではユーザが選択した入力データの正解クラスや予測結果，LIME の説明のバーチャートが表示されていて，図 4(右) には入力データを表示している．テキスト分類の例においては入力データは文書で，説明に用いられた特徴量 (単語) をハイライトしている．実際に入力データを観察することで，生成された説明の妥当性の分析ができる．

4. 適用事例

本章では，提案手法を糖尿病の分類タスクに適用した事例を紹介する．使用したデータセットは，Pima Indian Diabetes Dataset(PIDD)[Smith 88] である．これはピマインディアンという部族の 768 人の女性の健康状態に関するデータセットで，年齢や血中グルコース値，body mass index(BMI) など 8 つの特徴量で構成されている．タスクは positive(糖尿病) と negative(健康) の二値分類を行う．先行研究 [Ming 19][Cheng 20] にならい，データセットのうち 75%は訓練データ，25 %はテストデータとし，2 層のニューラルネットワークのモデルを訓練して精度 74%のモデルを得た．

本手法をこのモデルに適用し，LIME と Anchor のうち 1 つ

以上の説明がついた入力データの要約を図 1 に，ヒートマップを図 3 に示す．ヒートマップを見ると，グルコースの値が高いとき ($glucose > 141$) に赤い重みがついていて，血中グルコース値が高いと糖尿病の可能性が高いという，医学的な知見と一致する知識を正しく学習していることがわかる．一方でインスリン値が低い ($insulin <= 0$) ときに positive の重みがついており，これはローカルビューで入力データを確認すると欠損値であったため，取り除いて再学習する必要があることがわかった．

またこのタスクにおいて最も改善が必要とされるのは，病気の見逃しに繋がる false negative の部分であるため，このクラスタを詳細に観察する．Anchor の要約 (図 1(中)) の false negative のクラスタを見てみると，グルコース，年齢，BMI の低さなどを予測の根拠としていることがわかる．一方 LIME(図 1(上)) を見ると，negative な予測根拠としている特徴量も多数存在するものの，グルコースや年齢，BIM が高いことを positive な予測の根拠としている．ヒートマップの false negative のクラスタを見ると，クラスタの 1 つは，高グルコース ($glucose > 141$ に LIME の positive な重みがついている入力データで構成されている．他のクラスタでは，高グルコース ($glucose > 118$) の場合は LIME の positive な重みがついていて，低グルコース ($glucose < 118$) の場合は LIME も Anchor も negative の根拠としていることがわかる．グルコースに関して正しい知識を学習しているものの，Anchor の説明に出ている年齢や BIM の低さ，血圧の影響で予測に失敗していることが推測できる．より信頼できるモデルを作るには，グルコース以外の特徴量をより適切に学習するように特徴量エンジニアリングやデータクリーニングを行う必要がある．

LIME と Anchor の 2 つの説明手法を採用し大規模な可視化をしたことで，両方の説明を比較しながらエラー分析を進め，モデルが正しく学習している知識と改善が必要な点を効率よく理解可能になった．

5. おわりに

本論文では複数の説明手法を用いて機械学習モデルを解釈・改善するための可視化分析を提案した．複数の説明手法の比較によって，モデルの改善点を効率的に分析することが可能になった．今後の課題としては，説明のクラスタリングやフィルタリングを行いながらモデルの探索を実施できるように，よりインタラクティブな設計にしていくことや，実際にモデル改善まで実施して本手法の有効性を検証することが考えられる．

謝辞

本研究は JST CREST JPMJCR19A4，および，JSPS 科研費 JP21H03445 の支援を受けたものです．

参考文献

[Cheng 20] Cheng, F., Ming, Y., and Qu, H.: DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models (2020)

[Koh 17] Koh, P. W. and Liang, P.: Understanding Black-box Predictions via Influence Functions, in *International Conference on Machine Learning*, pp. 1885–1894 (2017)

[Krause 17] Krause, J., Dasgupta, A., Swartz, J., Aphinyanaphongs, Y., and Bertini, E.: A Workflow for Visual Diagnostics of Binary Classifiers Using Instance-Level Explanations, in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 162–172, IEEE (2017)

[Lundberg] Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. eds., *Advances in Neural Information Processing Systems 30*, pp. 4765–4774, Curran Associates, Inc.

[Ming 19] Ming, Y., Qu, H., and Bertini, E.: RuleMatrix: Visualizing and Understanding Classifiers with Rules, Vol. 25, No. 1, pp. 342–352 (2019)

[Ribeiro 16] Ribeiro, M. T., Singh, S., and Guestrin, C.: "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*, KDD '16, pp. 1135–1144, ACM (2016)

[Ribeiro 18] Ribeiro, M. T., Singh, S., and Guestrin, C.: Anchors: High-Precision Model-Agnostic Explanations, in *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)

[Shi 00] Shi, J. and Malik, J.: Normalized Cuts and Image Segmentation, Vol. 22, No. 8, pp. 888–905 (2000)

[Smith 88] Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R.: Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus, pp. 261–265 (1988)

[Spinner] Spinner, T., Schlegel, U., Schäfer, H., and El-Assady, M.: explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning, pp. 1–1

[Wachter 17] Wachter, S., Mittelstadt, B., and Russell, C.: Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, Vol. 31, No. 2, pp. 841–888 (2017)

[Wexler 20] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J.: The What-If Tool: Interactive Probing of Machine Learning Models, Vol. 26, No. 1, pp. 56–65 (2020)

[Zhang 19] Zhang, J., Wang, Y., Molino, P., Li, L., and Ebert, D. S.: Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models, Vol. 25, No. 1, pp. 364–373 (2019)