The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

特徴パターンを用いた機械学習の説明手法

Model-agnostic Explainer using Feature Patterns

浅野孝平 *1　全眞嬉 *1　徳山豪 *1
Kohei Asano　Jinhee Chun　Takeshi Tokuyama

*1東北大学情報科学研究科
Graduate School of Information Sciences, Tohoku University

Recently, high performance, though very complex, machine learning models have been proposed. Being able to interpret such black boxed decision making models is clearly critical for sensible tasks. In this paper, we propose a model-agnostic explanation method that details the prediction behavior of any models using feature patterns. Since the relationship between features is still unclear in the previous works, we focus on a combination of these features. We propose an algorithm which finds several minimal feature patterns that lead target prediction sufficiently using hill climbing search. In our experiments, we aim at measuring the faithfulness of our explanation, thus apply our method to sentiment analysis dataset and evaluate the faithfulness using two metrics: recall and precision. At last, we demonstrate the benefit of our method based on some image classification use cases with a black-box model.

1. はじめに

近年，深層学習をはじめとする高い識別性能をもつ機械学習モデルが様々な分野に応用されている．それらのモデルの多くはブラックボックスであり，ユーザがモデルの挙動や予測の原因について知ることが困難になっており，モデルや予測結果に解釈性に関する研究が活発に行われている．予測の原因となる特徴を特定する手法として Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro 16] がある．LIME では，個々の特徴の重要性を測ることはできるが，重要な特徴の組み合わせを特定することはできない．そこで，本研究では予測に影響を与えた特徴の組み合わせに着目した，新たなモデル依存性のない説明手法を提案する．提案手法では，特徴の組み合わせの中から，対象となる予測結果を得るために必要な特徴の組み合わせを探索アルゴリズムによって発見する．計算機実験によって提案手法の説明能力を定量的に評価した．また，画像分類への応用を行い，提案手法の有用性を検証した．

2. Local Interpretable Model-agnostic Explanations (LIME)

識別モデル $f : \mathcal{X} \to \mathbb{R}$ から得られた，インスタンス $x \in \mathcal{X}$ の予測結果 $f(x)$ に対して解釈を与えることを考える．ここで，x を被説明データと呼び，$\mathcal{X}$ は x のドメインである．LIME では，はじめに被説明インスタンス x をバイナリベクトル $\mathbf{x} \in \{0,1\}^d$ に変換する．例えばインスタンス x がカラー画像である場合，$\mathcal{X}$ は階数 3 のテンソルである．ここで，画像をピクセルやスーパーピクセルなどの手法を用いて領域分割し，その領域の有無で画像を表現することで，x をバイナリベクトルとして表現できる．

LIME ではスパースな線形関数 $g(x) = \mathbf{w}^\top \mathbf{x}$ によって，f を x の近傍で局所的に近似する．重みベクトル $\mathbf{w}$ から，識別に大きな影響を及ぼす $\mathbf{x}$ の特徴が特定できるため，予測結果に解釈を与えられる．g のように，予測結果に解釈を与えるモデルを説明モデルと呼ぶ．

次に，g の生成法について述べる．はじめに $\mathbf{x}$ の近傍のデータを $\mathbf{x}$ 中の 1 となっている特徴量をランダムに 0 にすることでサンプルし，これをサンプルベクトルと呼ぶ．次に，N 個のサンプルベクトル $\mathbf{x}_k$ $(k = 1, \ldots, N)$ を生成し，それらを元のデータ表現 $x_i \in \mathcal{X}$ に戻し，$\{(\mathbf{x}_k,\ \pi(x_k)f(x_k)) : k = 1, \ldots, N\}$ を訓練データとして Lasso 回帰することで，g を導出する．ここで，π は類似度関数であり，これを用いて局所性を測る．

以下では，この線形モデルによる定式化を Linear-LIME と記す．

連絡先: asano/jinhee/tokuyama@dais.is.tohoku.ac.jp

3. 極小パターンを用いた LIME の定式化

本論文では，極小な特徴パターンを用いた LIME (Minimal Patterns-LIME: MP-LIME) の定式化を提案する．MP-LIME では $\mathbf{x}$ の非零の要素を特徴の集合とみなし，$\mathcal{I} = \{1, \ldots, d\}$ と記す．また，特徴の部分集合 $e \in 2^{\mathcal{I}}$ を特徴パターンと呼び，ある特徴パターン e に対応するもとのデータ表現を x_e と記す．MP-LIME では説明モデル $\mathcal{E}_{\exp}$ を極小な特徴パターンの集合として定式化する．極小な特徴パターンを定義 1 として定める

定義 1.

$$f(x_{e_{\min}}) \sim f(x) \tag{1}$$

$$\forall i \in e_{\min},\ f(x_{e_{\min}\setminus\{i\}}) \nsim f(x) \tag{2}$$

を満たす $e_{\min}$ を極小な特徴パターンとして定義する．ここで，$\sim$ は右辺と左辺の識別結果が同じクラスに属することを表す比較演算子である．

定義 1 では，極小な特徴パターンを被説明インスタンス x と同じクラスに属するために最低限必要な特徴の集合として定義している．この定式化では，説明モデルに含まれるパターンは必ず被説明インスタンス x と同じクラスに属することが保証されており，識別に重要な特徴の組み合わせを表現できると考えられる．

次に，極小な特徴パターンの探索アルゴリズムについて述べる．探索はモデル f の評価値に基づいた山登り探索によって行う．はじめに $\mathcal{I}$ を初期状態として，$\mathcal{I}$ の近傍の特徴パターンすなわち，全ての $i \in \mathcal{I}$ について $f(x_{\mathcal{I}\setminus\{i\}})$ を評価する．こ

のとき，最も評価値の高いパターンを次の状態として，同様に近傍のパターンを評価して，極小な特徴パターンを探索する．この探索法で複数の極小な特徴パターンを発見するためには，再度繰り返し探索が必要となる．そこで，一度評価したパターン e のうち $f(x) \sim f(x_e)$ を満たすパターンは極小パターンの候補として $\mathcal{E}_{\text{cand}}$ に保存し，パターンを評価するたびに更新する．ひとつの極小な特徴パターンを発見したら，

$$\forall e_{\min} \in \mathcal{E}_{\text{exp}},\ e_{\min} \cap e \neq \varnothing \tag{3}$$

を満たすパターン e を $\mathcal{E}_{\text{cand}}$ から削除し，縮約された $\mathcal{E}_{\text{cand}}$ の要素の中から最も評価値の高いパターンを次の状態として再度探索を行う．式 (3) の制約を加えて探索を行うことで，評価されるパターン候補を削減する．

説明モデルの構築において，全ての極小な特徴パターンの列挙すると，評価されうるパターンの組み合わせは 2^d 個あるため，計算時間が膨大になりうる．そこで，極小な特徴パターンを L 個の発見するまで探索を行うことによって，この問題を回避する．ひとつの極小な特徴パターンを発見するために必要な評価回数は $\mathcal{O}(d^2)$ なので，L 個発見する場合でも $\mathcal{O}(Ld^2)$ の評価回数である，全パターンを探索するより効率が良い．

4. 計算機実験

4.1 説明能力の定量的評価

LIME と MP-LIME の説明能力を計算機実験によって評価した．説明能力は，識別モデルがデータを識別する際に使用した重要な特徴を LIME 及び MP-LIME によって予測した．本実験では識別モデル f として，ホワイトボックスなモデルである決定木を用いた．データセットとして，レビューデータ (books, DVD) [Blitzer 07] を用いた．各レビューは Word one-hot 表現で数値化し，1600 サンプルを訓練データとし，400 サンプルをテストデータとして用いた．Linear-LIME のパラメータは $N = 15000$ とし，正の重みを持つ特徴のうち上位 10 個を予測結果とした．MP-LIME については $L = 3$ とした．

説明能力を Precision, Recall を用いて評価した．表 1 に結果を示す．表 1 から，MP-LIME は Linear-LIME に比べ Recall は低くく，Precision が高いことが確認された．Precision は予測した特徴のうち，実際に予測に用いられた特徴の割合を意味するため，Precision が低い場合，予測した特徴の多くは識別結果に影響を与えていないことを意味している．そのため，MP-LIME の予測した特徴の多くが識別に影響したものであるため，Linear-LIME に比べ高い説明能力を持つと考えられる．

表 1: 説明能力の比較結果

	Recall		Precision	
	books	DVD	books	DVD
Linear-LIME	0.932	0.948	0.134	0.202
MP-LIME	0.768	0.758	0.980	0.980

4.2 画像分類への応用

深層学習モデルを用いた画像識別への応用を行う．Google が公開しているモデルである Inception-v3[*1] に図 1 を入力すると，96.8% で goose であると識別された．Inception-v3 はブラックボックスな畳み込みニューラルネットワークであり，このモデルが識別の際に画像のどの領域を重要視しているかは不明である．はじめに，画像をスーパーピクセルによって 40 分割し，画像を二値表現 (図 2) した．図 3 に $N = 10000$ とし，上位 10 個のスーパーピクセルについての Linear-LIME の出力結果を示す．また，図 4 に MP-LIME から得られたふたつの極小パターン ($L = 2$) を示す．赤色と青色でマスクしたスーパーピクセルがそれぞれ極小パターンを表している．

Linear-LIME による説明結果 (図 3) からは重要なスーパーピクセルは分かるが，重要な組み合わせは分からない．図 4 では，2 つのガチョウの頭部がそれぞれ説明結果として独立して出力されている．このことから，MP-LIME では，複数の特徴パターンが識別結果に影響している場合に，それらを検知できると考えられる．

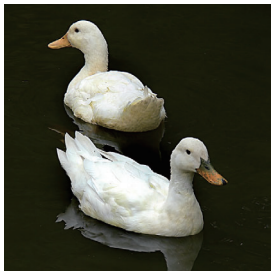

図 1: 元画像

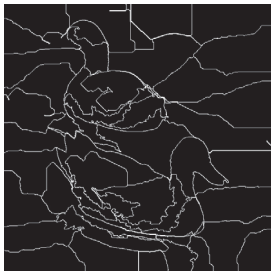

図 2: スーパーピクセルによる領域分割

図 3: Linear-LIME の出力

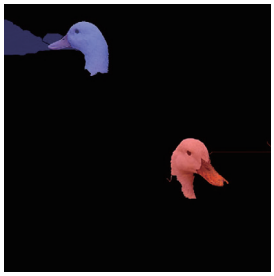

図 4: MP-LIME の出力

5. まとめ

本研究では，極小パターンの集合を説明モデルとした LIME の定式化である MP-LIME を提案した．計算機実験では，従来の Linear-LIME よりも高い Precision で説明が可能であることが示された．また，画像分類への応用より，複数の特徴の組み合わせが識別結果に影響している場合，それらを特定できることが確認された．

参考文献

[Blitzer 07] Blitzer, J., Dredze, M., and Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, in *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447 (2007)

[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*, pp. 1135–1144ACM (2016)

∗1 https://github.com/tensorflow/models