F-040

2分決定図を用いた機械学習予測の説明手法
Explanation Method of Machine Learning Prediction using Binary Decision Diagram

○浅野 孝平 *　　全 眞嬉 *　　徳山 豪 *
Kohei Asano　　Jinhee Chun　　Takeshi Tokuyama

1　はじめに

近年 Deep Neural Network (DNN) などをはじめとする高性能な識別モデルが提案されている．しかしながら，それらのモデルの多くはブラックボックスであるため，そこから得られた予測結果の原因が分からないという問題が生じている．ここで，判定された原因を特定することを予測結果に解釈を与えると呼ぶ．これは医療などの予測結果の解釈が求められる分野に機械学習技術を応用する際の重要な課題である．

任意の識別モデルから得られた予測結果に解釈を与える手法として, Ribeiro らは Local Interpretable Model-agnostic Explanations (LIME)[1] を提案した．本論文では，識別結果に寄与する特徴のパターンを 2 分決定図で表現する LIME を提案する．本手法では，説明モデルを特徴のパターンの集合とすることで，2 分決定図による識別パターンの可視化を実現する．そして，計算機実験によって提案手法の有効性を検証した．

2　LIME

識別モデル $f : \mathcal{X} \to \mathbb{R}$ から得られた，データ $x \in \mathcal{X}$ 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，$\mathcal{X}$ はデータのドメインである．LIME では，はじめに被説明データ x をバイナリベクトル $\mathbf{x}' \in \{0,1\}^d$ に変換する．これは，元のデータ表現の重要な成分が必ずしも人間にとって理解できるものでないためである．例えばデータ 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 $\mathbf{x}'$ は，画像をピクセルやスーパーピクセル [2] の有無で表現される．

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 $\mathbf{z}' \in \{0,1\}^d$ 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する．ここで $\mathbf{z}'$ をサンプルデータと呼ぶ．n 個のサンプルデータ $\mathbf{z}'_i$ $(i = 1, \ldots, n)$ を生成し，それらを元のデータ表現 $z_i \in \mathcal{X}$ に戻し，$\{(\mathbf{z}'_i,\ \pi_x(z_i) f(z_i)) : i = 1, \ldots, n\}$ 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_x は類似度関数であり，これを用いて局所性を測る．

3　2 分決定図を用いた LIME

本論文では，識別パターンを 2 分決定図によって可視化できる LIME (decision diagram LIME: ddLIME) を提案する．ddLIME では，説明モデルを重要かつ頻出な特徴の組み合わせパターンの集合とすることで 2 分決定図によって可視化が実現できる．

はじめに，頻出パターンの集合の定義について述べる．$\mathbf{x}'$ の特徴 $\{1, \ldots, d\}$ をアイテム集合 $\mathcal{I}$ とみなし, サンプルデータ $\mathbf{z}'$ の非零の特徴の集合 $\left\{j : z'_j = 1\right\} \subset \mathcal{I}$ をトランザクション t とみなす．そして，x と同じクラスに属するサンプルデータ $\mathbf{z}'$ に対応するトランザクションの集合をデータベース $\mathcal{T}$ とみなす．ここで，同じクラスに属することを $f(x) \sim f(z)$ で表すと，データベース $\mathcal{T}$ は，$\mathcal{T} = \{t_i : f(x) \sim f(z_i),\ i \in \{1, \ldots, n\}\}$ と表せる．あるパターン $P \subset \mathcal{I}$ の頻度を $freq(P)$ と表し $freq(P) = \#\{t : P \subset t \in \mathcal{T}\}$ で定義される．そして，最小サポート値と呼ばれるパラメータを $\theta \in \mathbb{N}$ とすると，頻出パターンとは $freq(P) \geq \theta$ を満たすパターン P である．したがって頻出パターンの集合は，$\{P : freq(P) \geq \theta\}$ で定義される．

次に, 各特徴の重要性について述べる．アイテム $i \in \mathcal{I}$ の重要性 p_i を以下のように定義する．

$$p_i = freq_p(\{i\}) - freq_n(\{i\}) \qquad (1)$$

ここで，$freq_p(\{i\})$, $freq_n(\{i\})$ はそれぞれデータベー

*東北大学大学院 情報科学研究科

ス $\{t : f(x) \sim f(z)\}$, $\{t : f(x) \nsim f(z)\}$ における $\{i\}$ の頻度である．ddLIME では上位 K 個の重要な特徴のみを考慮する．アイテム集合を式 (1) の重要性に基づいて $\mathcal{I}' \subset \mathcal{I}\ (\#\mathcal{I}' = K)$ に削減する．そして，データベース $\mathcal{T}$ を $\mathcal{I}'$ の要素からのみ構成されるように縮約する．

以上の定義を用いて ddLIME の説明モデルは以下のように定義する．

$$\mathcal{F} = \{P \in \mathcal{T}' :\ freq(P) \geq \theta\} \qquad (2)$$

$P \in \mathcal{T}'$ の頻度は最悪でも $\mathcal{O}(nd)$ の計算量で求められる．そして，$\mathcal{F}$ は Minato が提案した Zero-suppressed binary Decision Diagram(ZDD)[3] を用いて 2 分決定図として表現できる．

4 実験

4.1 説明能力の評価

LIME と ddLIME 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 ddLIME によって予測した．本実験では識別モデル f として，線形モデルと決定木を用いた．データセットとして，レビューデータ (books, DVD) [4] 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．LIME, ddLIME ともにサンプルデータ数は $n = 15000$，説明長は $K = 10$ とした．

説明能力は識別モデルが線形モデル (LM) であるとき，Precision を用い，決定木 (DT) のときは Recall を用いて評価した．Table 1 に結果を示す．Table 1 から，LIME と ddLIME はおおよそ同等の説明能力を持っていることが確認できる．この結果から式 (1) で定義した重要性が予測結果に解釈を与える際に有効であることが示された．

Table 1: Comparision of the precision and recall

	Precision (f: LM)		Recall (f: DT)	
	books	DVD	books	DVD
LIME	0.838	0.843	0.871	0.909
ddLIME	0.826	0.827	0.861	0.900

4.2 DNN による画像分類への応用

ブラックボックスな識別器である DNN から得られた識別結果に ddLIME を用いて解釈を与える．Google が公開している DNN モデルである inception-v3 に Fig. 1 に示す画像を入力したところ，26.5% で ”golden retriever” , 5.8% で ”tabby cat” であると識別された．はじめに，画像をスーパーピクセル [2] によって 35 分割し，画像をバイナリベクトルとして表現した．サンプルデータ数を $n = 4000$ として，各結果に解釈を与えた．

式 (1) の重要性に基づいて，”golden retriever” の識別に重要なスーパーピクセルを 8 個求め，その結果を Fig. 2 に示す．これより，”golden retriever” に対応するスーパーピクセルを提示できていることが確認できる．

次に，$K = 4$, $\theta = 100$ としたときの”tabby cat”の説明モデルを表す ZDD を Fig. 3 に示す．これより，”tabby cat”の識別に寄与している特徴のパターンが，2 分決定図で可視化できていることが確認できる．

Fig. 1: The original image

Fig. 2: Superpixels represent ”golden retriever”

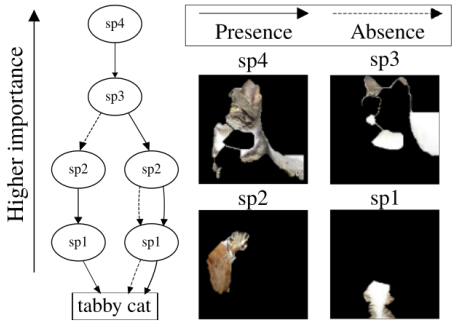

Fig. 3: Explanation model of ”tabby cat”

5 まとめ

本論文では，ddLIME を提案し，識別パターンの可視化を実現した．

謝辞

本研究は，総合科学技術・イノベーション会議が主導する革新的研究開発推進プログラム (ImPACT) の一環として実施したものです．

参考文献

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. ACM, 2016.

[2] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. Springer, 2012.

[3] Shin-ichi Minato. Zero-suppressed bdds for set manipulation in combinatorial problems. ACM, 1993.

[4] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. 2007.