F-023

予測の不確実性を考慮するデータ選択に基づくベイズ線形回帰のためのプールベース能動学習

Pool-based Active Learning for Bayesian Linear Regression Based on Uncertainty Sampling

永濱 僚基 † 内村 優太 † 菅谷 信介 ‡ 渡會 恭平 ‡ 鈴木 英之進 §
Tomoki Nagahama Yuta Uchimura Shinsuke Sugaya Kyohei Watarai Einoshin Suzuki

1. はじめに

能動学習 [1] は，教師付き学習においてできるだけ少ない訓練データで正確な予測を行うことを目的とした機械学習の手法の一つである．能動学習では，アルゴリズムが教師信号無しデータのうち学習に有益なデータを選択あるいは生成し，そのデータに対してオラクルと呼ばれる人間または機械が教師信号を付与するという過程を繰り返す．このうちプールベース能動学習 [1] は，教師無しデータを一度に入手した後，データプールの中から比較的少数のデータに教師信号を付与して訓練データに加える．

能動学習のための追加訓練データを選択するアルゴリズムのうち，uncertainty sampling[2] は，予測結果の不確実が高い教師信号無しデータを選択する手法である．簡潔で多くの分類・回帰学習の問題に適用できる一方で，回帰学習においては予測の不確実性は自明ではなく，目的変数の分散などを用いて間接的に低減する方策がとられる．本研究で扱うベイズ線形回帰では，目的変数の予測に加えてその予測結果の信頼度の評価を行うことができ，これを不確実性の指標として用いる．

実世界問題では教師付きデータの取得にコストがかかる場合があり，限られた教師付きデータの中で能動学習が効果的に行われるか評価することは重要である．ベイズ線形回帰を用いた不確実性の推定方法が，特に訓練データ数が極めて少ない場合に汎化誤差に与える影響を，他の不確実性の指標と比較しながら実験で評価する．

2. 能動学習手法

倉田ら [3] は，ベイズ線形回帰によって得られる予測分布の分散を用いて不確実性の推定を行っている．2.1 節では [3] に基づくプールベース能動学習手法を紹介し，2.2 節でその比較手法について述べる．

2.1. ベイズ線形回帰のためのプールベース能動学習

(0) 回帰モデルの設定

与えられた K 次元ベクトル $\mathbf{x} = (x_1, \ldots, x_K)$ を説明変数として，スカラー値である目的変数 y を推定する時，重みパラメータ $\mathbf{w}$ を用いて

$$\hat{y} = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) \tag{1}$$

と設定する．ただし，$\phi(\mathbf{x})$ は基底関数であり，ここでは $\phi(\mathbf{x}) = (1, x_1, \ldots, x_K)$ の線形基底を用いる．

(1) 初期訓練

まず，目的変数 y が平均 μ，分散 β^{-1} の正規分布 $\mathcal{N}(y|\mu, \beta^{-1})$ に従うとする．

$$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(y|\mu, \beta^{-1}) \tag{2}$$

次に重みパラメータ $\mathbf{w}$ の事前分布について平均を 0，分散共分散行列を単位行列 $\mathbf{I}$ を用いて $\alpha^{-1}\mathbf{I}$ とし，正規分布 $\mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$ に従うと仮定する．

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{3}$$

$\mathbf{w}$ の事前分布として共役な正規分布を選ぶため，事後分布も正規分布となる．スコア無しデータプール $\mathcal{U}_0$ から初期訓練データ $\mathcal{L}_1 = (\mathbf{x}_1, y_1)$ をランダムに選んで与えるときの $\mathbf{w}$ に関する事後分布は，y についての尤度関数 $p(y|\mathcal{L}_1, \mathbf{w})$ と $\mathbf{w}$ の事前分布から求められ，その平均 $\mathbf{m}_1$ と分散共分散行列 $\mathbf{S}_1$ とともに以下で表される．

$$p(\mathbf{w}|\mathcal{L}_1) = \mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{S}_1) \tag{4}$$

$$\mathbf{m}_1 = \beta \mathbf{S}_1 \mathbf{\Phi}_1^{\mathrm{T}} \mathbf{y}_1 \tag{5}$$

$$\mathbf{S}_1^{-1} = \alpha^{-1}\mathbf{I} + \beta \mathbf{\Phi}_1^{\mathrm{T}} \mathbf{\Phi}_1 \tag{6}$$

ここで，$\mathbf{\Phi}_n$ はその i 行目を $(1, x_{i_1}, \ldots, x_{i_K})$，$i = 1, \ldots, n$ とする計画行列である．$\mathbf{w}$ について事後分布を最大化するのは平均 $\mathbf{m}_1$ であるから，$\mathbf{w} = \mathbf{m}_1$ としてパラメータを更新する．

(2) 追加訓練データの選択

スコア無しデータ $\mathbf{x}_\mathcal{U}$ の学習における有益さを示す関数 $\delta(\mathbf{x}_\mathcal{U})$ を導入する．$N(N > 1)$ 回目の訓練の場合，$\delta(\mathbf{x}_\mathcal{U})$ の値を最大化するスコア無しデータ $\mathbf{x}_\mathcal{U}^*$ にスコアリングを行い，訓練データ $\mathcal{L}_N$ に追加する．

$$\mathbf{x}_\mathcal{U}^* = \arg\max_{\mathbf{x}_\mathcal{U}} \delta(\mathbf{x}_\mathcal{U}) \tag{7}$$

$$\mathcal{L}_N = \mathcal{U}_{N-1} \cup \mathbf{x}_\mathcal{U}^* \tag{8}$$

$$\mathcal{U}_N = \mathcal{U}_{N-1} \setminus \mathbf{x}_\mathcal{U}^* \tag{9}$$

$\delta(\mathbf{x}_\mathcal{U})$ は，目的変数についての予測分布から推定する．目的変数が正規分布に従うとすると，スコア無しデータ $\mathbf{x}_\mathcal{U}$ の目的変数 y についての予測分布は

$$p(y|\mathbf{x}_\mathcal{U}, \mathcal{L}_N) = N(y|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_\mathcal{U}), \sigma^2(\mathbf{x}_\mathcal{U})) \tag{10}$$

$$\sigma^2(\mathbf{x}_\mathcal{U}) = \beta^{-1} + \phi(\mathbf{x}_\mathcal{U})^{\mathrm{T}} \mathbf{S}_N \phi(\mathbf{x}_\mathcal{U}) \tag{11}$$

†九州大学大学院 システム情報科学府 Kyushu University, Graduate School of Informaiton Science and Electrical Engineering
‡株式会社ビズリーチ BizReach,Inc.
§九州大学大学院 システム情報科学研究院情報学部門 Kyushu University, Faculty of Informaiton Science and Electrical Engineering

として表される．ここで，(3) 式は予測分布の分散であり，第 1 項は目的変数についてのノイズを示す．第 2 項の $\mathbf{S}_N$ は訓練データを N 個与えたときの $\mathbf{w}$ の事後分布における分散共分散行列であり，第 2 項が $\mathbf{w}$ に関する不確実さを表すため，$\delta(\mathbf{x}_\mathcal{U}) = \sigma^2(\mathbf{x}_\mathcal{U})$ とおける．分散を最大化するスコア無しデータ $\mathbf{x}_\mathcal{U}$ を選択することで，予測の不確実さが大きく低減されると考えられる．（本論文ではこのデータ選択手法を variance と呼ぶ.）

(3) 追加訓練

$N(N > 1)$ 回目の訓練の場合，$\mathbf{w}$ についての事後分布の平均 $\mathbf{m}_N$ と分散共分散行列 $\mathbf{S}_N$ は，次の式で表される．

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_{N-1}^{-1}\mathbf{m}_{N-1} + \beta\mathbf{\Phi}_N^{\mathrm{T}}\mathbf{y}_N) \quad (12)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_{N-1}^{-1} + \beta\mathbf{\Phi}_N^{\mathrm{T}}\mathbf{\Phi}_N \quad (13)$$

$\mathbf{w} = \mathbf{m}_N$ としてパラメータを更新する．訓練データが既定の数になるまで (2)，(3) を繰り返す．

2.2. 比較手法

[3] における実験では，訓練データ数が数百件における汎化性能を評価している．モデルを効率的に更新するために不確実性を推定して追加データを選択するが，不確実性の推定にもモデルの更新が必要であるから，訓練データが数件から数十件程である場合，数百件の場合と同じ性能が得られないおそれがある．そこで本論文では，1 件の訓練データで学習した後，アルゴリズムにより 1 件ずつ訓練データを追加する場合の汎化誤差の推移を評価する．

2.1 節 (2) のデータ選択アルゴリズムを評価するために，ベイズ線形回帰に関わらず回帰学習一般に適用可能なアルゴリズムを用いて比較を行う．D. Wu[4] らは “周辺に訓練データが存在しないスコア無しデータはその予測スコアの不確実性が高い” という考えに基づき，スコア無しデータの訓練データとの最小距離を求めることにより追加訓練データを選択する．

訓練データが N 個あるとき，スコア無しデータ $\mathbf{x}_\mathcal{U}$ について説明変数空間での訓練データ $\mathbf{x}_\mathcal{L}$ とのユークリッド距離の最小値 $\mathrm{d_x}(\mathbf{x}_\mathcal{U})$ を求める．

$$\mathrm{d_x}(\mathbf{x}_\mathcal{U}) = \mathrm{min}_i\, ||\mathbf{x}_\mathcal{U} - \mathbf{x}_{\mathcal{L}_i}||, \quad i = 1, ..., N \quad (14)$$

説明変数のみを推定に用いるため，モデルの更新を必要とせず実行が簡単であるが，重みの大小に関わらず全ての説明変数方向の距離を同等に扱っている．予測における説明変数ごとの重要度を考慮するため，目的変数空間での距離を推定する．目的変数空間においてスコア無しデータ $\mathbf{x}_\mathcal{U}$ に対する予測値 ${}^{\mathrm{T}}\mathbf{w}\phi(\mathbf{x}_\mathcal{U})$ と訓練データ $y_\mathcal{L}$ のユークリッド距離の最小値 $\mathrm{d_y}(\mathbf{x}_\mathcal{U})$ は，

$$\mathrm{d_y}(\mathbf{x}_\mathcal{U}) = \mathrm{min}_i\, ||\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_\mathcal{U}) - y_{\mathcal{L}_i}||,\ i = 1, ..., N \quad (15)$$

となる．$\delta(\mathbf{x}_\mathcal{U}) = \mathrm{d_x}(\mathbf{x}_\mathcal{U})$ または $\delta(\mathbf{x}_\mathcal{U}) = \mathrm{d_y}(\mathbf{x}_\mathcal{U})$ として，$\delta(\mathbf{x}_\mathcal{U})$ を最大化するスコア無しデータを選択する．（これらの手法をそれぞれ dist-x，dist-y と呼ぶ.）

3. 評価実験

Indeed 英語求人票データセット，Stanby 日本語求人票データセット，Boston 住宅価格データセットを用いて実験を行った．表 1 に Indeed 英語求人データセットにおける説明変数と目的変数，および事例の一部を示す．要求スキルやレビュー数などの複数の説明変数から目的変数を推定することを目的とする．

表 2 に各データセットについて実験で用いるデータサイズを示す．追加訓練データごとの平均 2 乗誤差の推移により，2.3 節で述べたデータ選択手法（variance，dist-x，dist-y）とランダムなデータ選択 (random) を比較評価した．初期訓練データは 1 個であり，スコアリングしたデータを 1 個ずつ加えて追加学習を行った．図 1 に Indeed 英語求人票においてテストデータをランダムサンプリングして上述の実験を 30 回行った平均値を示す．

最も汎化性能が優れていたのは dist-y であり，1〜20 回目までの訓練において，random より平均 2 乗誤差が平均 10.9% 減少した．一方，variance では平均 2 乗誤差が random より平均 6.0% 増加したが，約半分の訓練データで誤差が収束した．訓練データが極めて少ない段階では，パラメータの更新が進まずに不確実性の推定値自体も信頼性の低いものになったこと，また外れ値を選択したために汎化性能の向上に寄与しなかったことが，訓練初期に random より汎化誤差の減少が遅くなった原因として考えられる．

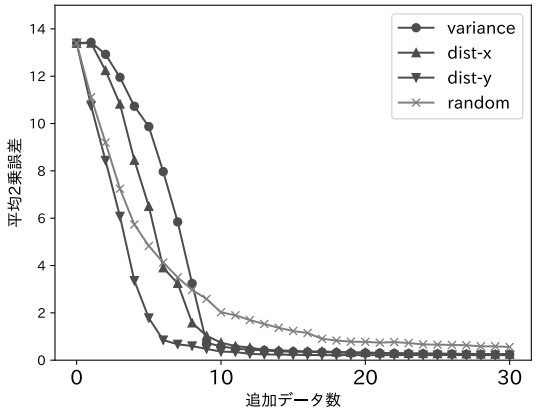

図 1: Indeed 英語求人票における追加訓練データ数ごとの平均 2 乗誤差

4. おわりに

本研究では，不確実性を考慮する訓練データ選択により，ランダムな選択と比較して汎化誤差を早く収束させる結果を得た．ただし，訓練初期においては，ランダム選択より汎化誤差の減少が遅い手法も存在した．解決法の一つして，データの代表性を推定することが挙げられる．代表性を考慮することにより，元のデータ分布に沿った事例を選択できる．予測の不確実性と

表 1: Indeed 英語求人票データセットの説明変数と目的変数

	変数	事例 1	事例 2	…
説明変数	要求スキル数	8	1	…
	レビュー数	484	16	
	求人投稿からの経過日数	30	52	
	求人テキストの単語数	363	121	
	職種ごとのキーワード数	2	0	
目的変数	レビュースコア	4.3	3.6	

表 2: データセットのサイズ

データセット	説明変数の数	目的変数の数	テスト事例数	スコア無し事例数
Indeed 英語求人票データセット	5	1	90	210
Stanby 日本語求人票データセット	10	4	30	70
Boston 住宅価格データセット	13	1	75	178

データの代表性の両方を考慮するデータ選択手法については，別稿で報告する．

参考文献

[1] B. Settles, “Active Learning Literature Survey”, Computer Sciences Technical Report, No.1648, University of Wisconsin–Madison (2009)

[2] D. D. Lewis, W. A. Gale, “A Sequential Algorithm for Training Text Classifiers”, In Proc. SIGIR ’94, page 3-12 (1994).

[3] 倉田 宗史，巽 啓司，谷野 哲三，平井 雄作，松岡 俊匡，谷 貞宏，“ベイズ線形回帰を用いた高精度逐次比較型 A/D 変換器の誤差補正のための追加学習法”，システム制御情報学会論文誌，29,(2)76-85 (2016)

[4] D. Wu, C. T. Lin, J. Huang, “Active Learning for Regression Using Greedy Sampling”, Information Sciences, 474, 90-105 (2019)