6M-08

ユーザの多様性を考慮したクラスタワイズ型機械学習モデルの提案とその解釈方法に関する研究

三橋 可奈 †　山下 遥 ‡　清水 良太郎 §
上智大学 †　上智大学 ‡　ZOZO研究所 §

1 はじめに

近年のメディアやデバイスの普及により，ソーシャルネットワークサービスを運営する企業はユーザの属性情報や投稿履歴データなどを蓄積し，活用できるようになった．データをもとに経営戦略などを立案するためには多様なユーザの属性や行動の特徴を適切に分析する必要がある．これらの多様性を考慮するために，ユーザに潜在的ないくつかの母集団を仮定して分析することが望ましい．これまで，データのグループ分け（クラスタリング）と重回帰分析を交互に行い，クラスとモデルを更新していくクラスタワイズ回帰分析 [1] が提案されている．さらに，クラスワイズ回帰分析により予測精度の高い機械学習手法の SVM を取り入れたクラスタワイズ SVM 分析法 [2] も提案されている．しかし，重回帰分析や SVM は外れ値などが存在するデータへの適用が難しいという問題がある．

そこで本研究では，重回帰分析や SVM と比較して予測精度が高く，外れ値などに頑健な XGBoost[3] をクラスタワイズのアルゴリズムに導入し，予測モデルの精度の向上を目指す．この時，2 つの問題点が存在する．1 つ目は，予測精度の高いモデルを導入することで予測に用いるデータ自体が学習に用いられたモデルに当てはまりすぎてしまい，初期のクラスから更新されない点である．それに対して本研究では，新たなデータの分割方法を提案する．2 つ目は，XGBoost を解釈する際に一般的に用いられる”Feature Importance”という値では十分にモデルを解釈できない点である．これを解決するため，SHAP[4] という機械学習モデルを解釈するための手法を導入する．提案手法を用いて実データを分析し，予測精度やクラスの更新率を従来手法と比較することで提案手法の有効性を示す．

2 先行研究

2.1 クラスタワイズ回帰分析

クラスタワイズ回帰分析は，データを K 個のクラスに分割し，クラスごとに回帰モデルを構築する方法である．サンプル数 N のデータに対するクラスタワイズ回帰分析の最適化問題は，i 番目のデータに対する p 個の変数からなる説明変数ベクトル $\boldsymbol{x_i} \in R^p$，目的変数 y_i，偏回帰係数ベクトル $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$，さらに i 番目のデータが k 番目のクラスに所属するか否かを表す変数 w_{ik} を用いて以下のように表される．

$$\underset{\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, w_{ik}\in\{0,1\}}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{i=1}^{N} w_{ik}(y_i - \hat{y}_{ik})^2$$

$$\text{s.t.} \sum_{k=1}^{K} w_{ik} = 1, \quad \hat{y}_{ik} = \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} \tag{1}$$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} = (\boldsymbol{X}_{\boldsymbol{k}}'\boldsymbol{X}_{\boldsymbol{k}})^{-1}\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{y}_{\boldsymbol{k}} \tag{2}$$

ただし，$\boldsymbol{X_k}$，$\boldsymbol{y_k}$ はクラス k に所属するデータ群に関する $\boldsymbol{x_i}$，および y_i を並べたものである．

2.2 XGBoost と SHAP

本研究で用いる XGBoost は，勾配ブースティングと呼ばれるアンサンブル学習と決定木を組み合わせた手法である．XGBoost を解釈する方法として Feature Importance を観察する方法がある．しかし，特徴量で連続値と離散値を扱う場合，Feature Importance を用いた方法では，適切な解釈が行えないという主張が存在する [4]．今回用いるデータには，連続値と離散値が含まれるため，モデルを解釈するためには他の評価手法が必要である．そこで本研究では，各特徴量の重要度を推定する手法として SHAP を導入する．この値は，各データに対する予測値と平均値の差分を各特徴量の貢献度によって分解した値であり，「なぜモデルはこのような予測を行なったのか」という予測の理由を解釈できる値として活用されている．

3 提案手法

3.1 提案モデル

本研究では，(2) 式の回帰部分を XGBoost のモデル推定に置き換えた新たなモデルを提案する．サンプルサイズ N のデータに対してクラス数を K，説明変数ベクトル $\boldsymbol{x_i}$，目的変数 y_i，各クラスのモデル f_k とし，以下で表す．

$$\underset{f_k, w_{ik}\in\{0,1\}}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{i=1}^{N} w_{ik}(y_i - \hat{y}_{ik})^2$$

$$\text{s.t.} \sum_{k=1}^{K} w_{ik} = 1, \quad \hat{y}_{ik} = f_k(\boldsymbol{x_i})$$

3.2 提案手法のアルゴリズム

従来のクラスタワイズは，クラスごとにモデルを学習し，目的変数の実測値から最も近い予測値を取るモデルのクラスへと所属クラスを更新していく手続きを繰り返すものである．このアルゴリズムに予測精度の高いモデルを導入した場合，予測に用いるデータ自体が学習に用いられたモデルに当てはまりすぎてしまい，初期のクラスから更新されない可能性がある（本研究ではこの問題を初期クラス依存とする）．そこで本研究では，データを K クラスに分けたのちにそれぞれのクラスのデータを L 個のサブクラスに分割し，学習データと予測データに振り分ける以下のようなアルゴリズムを提案する．

Step0　データ i が所属するクラスをランダムに決定する（$w_{ik}\in\{0,1\}$）

Step1　k 番目のクラスに入るデータ i の所属する l 番目のサブクラスをランダムに決定する（$w_{ikl}\in\{0,1\}$）．前の繰り返しで求めた w_{ik} を用いて最適なモデル f_{kl} を推定する．

$$\hat{f}_{kl} = \underset{f_{kl}}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{N} w_{ik}(1 - w_{ikl})(y_i - f_{kl}(\boldsymbol{x}_i))^2$$

Step2 Step1 で求めた $\hat{f}_{kl}$ を用いて以下の最適化を行う．

$$\underset{w_{ik}\in\{0,1\}}{\operatorname{argmin}} \quad w_{ik}w_{ikl}(y_i - \hat{f}_{kl}(\boldsymbol{x}_i))^2$$

Research on proposal of cluster-wise machine learning model considering user diversity and its interpretation approach
†Kana Mitsuhashi · Sophia University
‡Haruka Yamashita · Sophia University
§Ryotaro Shimizu · ZOZO Research

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

$$\text{s.t.} \sum_{k=1}^{K} w_{ik} = 1$$

Step3 Step1，2 を更新がなくなるまで繰り返す．

4 実データ分析

4.1 実験条件

本研究では，株式会社 ZOZO が運営するファッションコーディネートアプリ「WEAR」におけるユーザの属性データと投稿データを用いる．WEAR はユーザがコーディネート画像を投稿できるサービスであり，コメントや「いいね」といったリアクションができるという特徴を持つ．本研究では，直近１年の投稿の中でいいね数を多く獲得した上位 10 万件の投稿に絞り分析する．また，目的変数 y_i を「投稿がお気に入りされた回数」，性別などを含むその他 16 個の項目を説明変数 $\boldsymbol{x_i}$ として分析する．そして，事前分析により，クラス数を 4，サブクラスを 3 と設定する．

4.2 評価指標

提案手法により，予測精度が向上したのかを確かめるための指標として，平均絶対誤差（MAE）を用いる．

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y_i} - y_i|$$

また，初期クラス依存が解消されているかを確かめるためにクラスの更新率を用いる．更新率を以下のように表す．この値が小さいほどクラスが更新されたデータが少なく，初期に割り当てられたクラスから更新が十分に行われていないと考えられる．

$$\text{更新率} = \frac{\text{(元のクラス}\neq\text{更新されたクラス) となるデータ数}}{\text{全データ数}}$$

4.3 実験結果

提案手法を用いることで予測精度が向上したのかを確認するため，全データを１つの XGBoost で予測させた場合（従来 1）と比較する．また，初期クラス依存の問題が解決されているのかを確認するため，従来のクラスタワイズのアルゴリズムに XGBoost を導入した手法（従来 2）とも比較する．エポック数は従来 2，提案手法ともに 100 に設定した．

表 1 に各手法の MAE の結果を示す．

表 1: MAE の分析結果

手法	従来 1	従来 2	提案
MAE	22.6	9.20	8.94

以上の結果より，提案手法による予測精度の向上を示せた．次に従来 2 と提案手法の更新率を以下の表 2 で示す．

表 2: 更新率の分析結果

手法＼epoch	1	2	4	6	8	10	20
従来 2	0.73	0.48	0.26	0.15	0.12	0.10	0.05
提案	0.75	0.58	0.32	0.21	0.16	0.13	0.07

表 2 より epoch=2 の時，つまりクラスごとに１度モデルを学習したのち，初めてクラスタリングを行った際にクラスが変わったデータの割合が従来 2 では 0.48 であるのに対して提案手法では 0.58 である．つまり，提案手法の方がクラスが更新されているデータ数が多く，初期クラス依存が緩和されていることを示せた．

最後に，クラスごとのモデルに違いがあるのか，そして，モデルにおける特徴量の重要度がどのようなものかを確認するため SHAP 値を比較する．提案手法は各クラス 3 つずつモデルがあり，それに対応した SHAP 値が存在するが，各クラスのモデルの SHAP 値はクラスごとで大きな違いは無かった．今回は簡略化のためそれぞれの中から 1 つを抜粋して示す．

表 3: SHAP 値の比較

	従来 1	提案			
		cls1	cls2	cls3	cls4
性別	0.04	0.00	0.00	0.34	0.00
身長	0.72	0.35	0.43	0.53	0.77
年齢	1.4	0.91	0.52	0.19	0.60
WEARISTA か	0.25	0.00	0.00	0.00	0.00
スタッフか	1.95	1.18	1.10	0.81	2.08
フォロワー数	1.67	0.63	0.51	1.36	1.19
フォロウィー数	1.67	2.16	1.91	1.39	8.41
総投稿数	11.0	2.17	0.37	12.9	56.3
年数	1.16	0.79	0.09	0.04	0.15
投稿文字数	0.31	0.18	0.07	0.04	0.08
タグ数	4.45	4.96	1.01	6.82	1.25
投稿月	8.13	0.49	16.9	2.40	2.08
登録アイテム数	0.00	0.06	0.02	0.01	0.14
コメント数	6.22	1.58	1.05	3.05	14.6
閲覧数	13.2	9.12	4.03	14.0	10.4
いいね数	24.6	18.3	49.1	27.4	28.3

表 3 は各特徴量の SHAP 値の絶対値の平均であり，各特徴量の平均的な重要度として捉えることができる．クラス分割を行わない従来手法 1 は，いいね数，閲覧数，総投稿数の順で値が大きい．それに対して，提案手法のクラス 4 では投稿数，いいね数，コメント数の順であった．その他のクラスにおいても，クラスごとに順番や値の違いが見られた．以上の結果より，ユーザの多様性に合わせてモデルの推定，クラスタリングができたと考えられ，提案モデルの妥当性を示すことができたと考えられる．

5 おわりに

本研究では，多母集団を仮定したクラスタリングと回帰分析を同時に行う手法に予測精度の高い機械学習モデル XGBoost を導入する方法を提案し，そのアルゴリズムを示した．さらに，結果を解釈するための方法の適用を提案した．また，実データ分析を通してその妥当性に関して検討した．

今回のモデルは既存のデータから学習したものであり，現状では新たなデータに対しての予測を行うことは難しい．よって今後の課題としては，新たなデータを分類できるようなモデルの提案などが挙げられる．

参考文献

[1] Helmuth Späth. A fast algorithm for clusterwise linear regression. *Computing*, Vol. 29, No. 2, pp. 175–181, 1982.

[2] Kaisa Joki, Adil M Bagirov, Napsu Karmitsa, Marko M Mäkelä, and Sona Taheri. Clusterwise support vector linear regression. *European Journal of Operational Research*, Vol. 287, No. 1, pp. 19–35, 2020.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.

[4] 吉田秀穂, 田嶋優樹, 今井優作. 決定木モデルの解釈における shap 値の有用性の検証. 人工知能学会全国大会論文集 第 34 回全国大会 (2020), pp. 3E5GS204–3E5GS204. 一般社団法人人工知能学会, 2020.

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.