5C-03

変数の分布に着目した特徴量生成方法及びそれに適した機械学習方式の検討

高田　晋太郎†

(株)日立製作所　研究開発グループ　人工知能イノベーションセンタ†

1．はじめに

AI を用いたソリューションビジネスを展開するためには，機械学習による予測モデル生成が不可欠となりつつある．一般的に，高精度な予測モデルを得るためには，対象とする事象のデータ取得と，事象を表現する特徴量（説明変数）の設計，学習データの準備及び，その学習データに適した機械学習アルゴリズムの選定や，学習時における種々の条件（ハイパーパラメータなど）調整を必要とする．これらのうち，学習アルゴリズム選定や各種条件調整などは，近年では，モデル化対象の学習データに対して，人手を介することなく計算機側で網羅的に行う学習手段(AutoML)が存在する．一方で，現実の事象においては，収集したデータの各サンプルが，設計した説明変数上で様々な分布を持つこともあり，AutoML のような学習手段のみで所望の精度の予測モデルを得られないことがある．このような場合には，対象とする事象とデータに則した特徴量の設計と機械学習アルゴリズムの適用がより重要となる．

本研究では，このようにアルゴリズムの選定やパラメータ調整のみでは対応できないような分布を持つ学習データに対して，精度向上に貢献する特徴量の生成方法と，それに適した学習方式についての検討を行った．

2．予測精度向上のための一般的手段

データ収集と特徴量設計が済み，学習データが準備できている場合に，より良い精度の予測モデルを得るための一般的な手段について述べる．最も基本的な手段として，学習データと評価データの組み合わせを複数用意し，学習と条件調整を行い過学習を防ぐ交差検証法が挙げられる．また，複数のアルゴリズムで予測モデルを生成し，最も精度の高いものを選択することで，よりデータに則したアルゴリズムを決めることができる．更には，複数生成した予測モデルにおける予測値を統合して最終的な予測値とするアンサンブル学習も存在する．これは，予測値の平均をとるだけの単純な方法から，予測値を新たな特徴量とみなし，再度学習を行うStacking 法など，様々なものがある[1]．

本研究においては，これらの手段を用いても所望の予測精度に達しない場合，特に学習データのサンプルが複雑な分布を形成しているような場合でも，予測精度向上に有効な手段を確立することを目的とする．

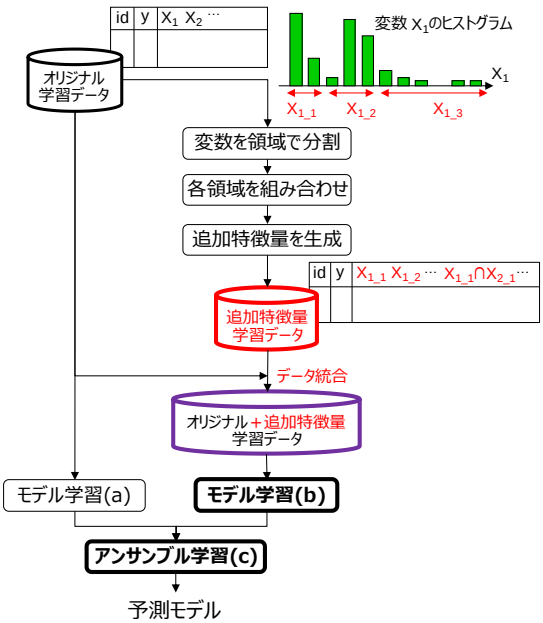

図１．検討した特徴量生成及び学習方式

3．特徴量生成と学習方式の検討

以下の観点に基づいて方式の検討を行った．

・予測精度向上に貢献する新たな特徴量を生成
・自動での特徴量生成実行が可能

前者については，新たな特徴量として明示的に生成し学習を行うことで，精度向上に寄与した要因(=特徴量)の把握が容易となり，対象事象の理解の助けとなる．後者については，自動で実行可能な精度向上手段として確立することで，

A study on the method of feature generation and machine learning focusing on the distribution of variables. †Shintaro Takada, †Hitachi, Ltd. Research & Development Group, Center for Technology Innovation – Artificial Intelligence.

より迅速なソリューションの展開が可能となる．

新たな特徴量の生成方法については様々な手段が考えられるが，機械学習アルゴリズムが学習データの学習の過程で，考慮しきれない条件等を表現できているものが望ましい．特に，現実の事象を対象としているデータにおいては，特定の値にサンプルが集中しているなど特殊な分布を持つことがある．これに対し，我々はこれまで特徴量の分布と目的変数との関係性に基づいた独自の特徴量生成手法を開発してきた知見を活用することを考えた[2]．

図 1 に，検討した特徴量生成及び，学習方式の概要を示す．

【特徴量生成方法】

学習データにおける各説明変数において，サンプルの分布に基づき，分割後の各領域に含まれるサンプル数が同じになるよう領域を定義する．さらに，異なる変数の領域同士の組み合わせも定義する．各サンプルが，定義された各領域に含まれるかどうかの二値(0/1)を持つ変数を特徴量として生成する．生成した特徴量のうち，目的変数と相関が高いものを追加特徴量とする．

【学習方式】

３通りの学習方式を定義し，上記生成した特徴量が最も活かされる方式について検討を行った．(a)：オリジナルの学習データでモデルを学習，(b)：オリジナルの学習データに生成した特徴量を追加して学習，(c)：(a)と(b)による学習をアンサンブルしたもの

これらの方式によって，自動で生成可能な特徴量の追加による予測精度の向上を期待する．

４．現実データに対するモデル精度評価

本方式の効果を確かめるため，現実の事象を対象としたデータに対し，予測モデルを生成し精度による評価を実施した．使用したデータは，ある物流倉庫内における荷物ピッキング作業時間の予測に関するものである．学習データのサンプル数は 14766 個，特徴量数は 104 個である．追加特徴量の生成にあたっては，各変数の領域を分割後のサンプル数が等しくなるよう 5 分割し，各変数間での組み合わせは 2 つまでを考慮した．生成した特徴量のうち，目的変数と相関の高い上位 100 個のものを追加特徴量とした．

表 1 に精度評価結果を示す．「モデル学習」の項目は，図１における各種学習方式(a)～(c)に対応する．使用した機械学習アルゴリズムは後ろに記載されており，LGBM は LightGBM，RF は Random Forest をそれぞれ使用した場合である．精度評価指標として MAE(Mean Absolute Error)と RMSE(Root Mean Squared Error)を用いた．

表１．精度評価結果

#	モデル学習	MAE	RMSE
1	(a):LGBM	112.48	177.84
2	(a):RF	113.70	178.12
3	**(b):LGBM**	**112.15**	**177.10**
4	(b):RF	113.88	178.24
5	**(c):(a)+(a)** (a:RF, a:LGBM)	**112.08**	**176.77**
6	**(c):(a)+(b)** (a:RF, b:LGBM)	**111.82**	**176.22**

一般的な学習方式である(a)について，精度の高い LGBM の結果がベースラインとなる(#1)．これに対し，特徴量を新たに追加したデータで学習した(b)において，LGBM による学習の場合(#3)に MAE，RMSE ともに#1 の場合よりも精度が向上していることが確認された．更に，(a)と(b)の学習のアンサンブルである(c)において(本評価では Elastic Net による Stacking を適用)は，(a)に RF，(b)に LGBM である組み合わせの場合(#6)に，最も高い精度を得られた．

本結果から得られる考察として，まず新たに生成した特徴量をオリジナルの学習データに追加することで，精度が向上する場合を確認した(#3)．これは，変数毎の分布に考慮して生成した特徴量の追加が精度向上に効果があったためと言える．また，(c)のアンサンブル学習は，同一の学習データに対して異なるアルゴリズムの学習を組み合わせる一般的な場合(#5)よりも，本方式の特徴量の追加によってより多様性を増した学習の組み合わせを行うことで(#6)，より高い精度を得られると言える．

５．まとめ

現実のデータを対象とした予測モデルの生成において，予測精度を高める手段の確立を目的とし，変数の分布に着目した特徴量生成と追加，及びアンサンブル学習の方式について検討した．精度評価の結果，既存の学習方法よりも高い精度を得られる見込みを得た．今後，精度向上要因の詳細な明確化と，特徴量生成方法の改良及び好適なユースケースの明確化を行っていく．

参考文献

[1]https://mlwave.com/kaggle-ensembling-guide/

[2]F. Kudo, et al.,“An Artificial Intelligence Computer System for Analysis of Social-Infrastructure Data” IEEE 17th Conference on Business Informatics, 2015.