

敵対的分類器を用いた再学習用データ選択手法の研究

理学専攻・情報科学コース 2140655 今野由麻

1. はじめに

機械学習モデル利用の課題として model decay がある。その原因として知られるデータドリフトに着目して、J. Pan らによる研究[1]を拡張した適切に選択した学習データを用いてモデルの更新を行う枠組み(データ選択手法)を提案してきた[2]。データ選択手法では、ドリフトしていないデータを判別するためにデータ選択指標が利用されている。本研究では、Adversarial Validation を用いた2つのデータ選択指標とその指標を用いたデータ選択手法について、人工ドリフトデータセットを利用した評価を行う。

2. 実験に利用した CIRCLES データセット

人工データセットである CIRCLES データセットは、M. Kubat ら[3]によって提案されたものである。本研究では、バッチ数が4でバッチ1つあたり 10,000 件のデータを含む CIRCLES データセットの設定を変えて複数生成し実験を行うが、バッチ1と特殊な終端バッチであるバッチ4は利用しない。CIRCLES データセット生成の際に用いた設定の説明を以下に示す。

- Transition Length(t1) : ドリフトによりデータが変質しきるまでの移行期間の長さ
- ドリフトパターン : 複数の円で構成されており、データ生成時に使用する

CIRCLES データセットの特徴量は0から1までの二次元空間上の座標であり、クラスは正か負の二値である。特徴量に対するクラスの決め方は、その座標がある円の内側にあれば負、外側にあれば正とする。また、クラスの割合はほぼ半数ずつ生成される。CIRCLES データセットのバッチ n (ただしバッチ n が終端の場合を除く)の生成には、複数の円で構成されたドリフトパターンのうちの円 n と円 n+1 を使用する。バッチ n に含まれるデータを時系列に一件ずつ生成する際、初めは円 n に基づいてクラスを決定するが、後半の t1 個分では、徐々に円 n+1 に基づいて生成されるデータの割合が増加していく。

表1には、実験に利用するデータセット生成のために利用したドリフトパターン original を示した。ドリフトパターン original は CIRCLES データセット提案者の M. Kubat ら[3]が利用したものである。

3. データ選択指標について

用意したデータ選択指標は、以下の二つである。その両方の指標において、学習に利用可能なデータ(バッチ 2)と予測したいデータ(バッチ 3)を見分けるように学習した敵対的分類器から得たクラス予測確率を利用する。本研究で利用する敵対的分類器や予測のためのモデルには、Gradient Boosted Decision Tree を利用した。

データ選択指標1 : 学習に利用可能なデータのうち予測したいデータのクラスへの予測確率が高いものを選択する

データ選択指標2 : 学習に利用可能なデータのうち予測したデータのクラスへの予測確率が50%付近のものを利用する

表2には、全部で10,000件のデータを含むバッチ2から、それぞれのデータ選択指標に基づいて2,000件/4,000件/6,000件/8,000件/10,000件(全件)を選択してバッチ3の予測を行なった結果を示す。実験には、ドリフトパターン original と3つの t1 (2,000/6,000/10,000) を用いて生成したデータセットを利用した。掲載した結果は、シードを変えて30回実験を行った Area Under the Curve (以下、AUC と称する) スコアの平均値であり、各行の最良の値を太字で表記した。

表2の結果では、t1 が短い場合のみ指標2より指標1でよい結果が得られた。しかし、t1 が短く学習に有用なデータが少ない環境であるため、指標1による性能改善の幅は4.5%ほどである。t1 が大きい場合では、指標2を用いた場合の性能改善の幅は8.4%ほどである。

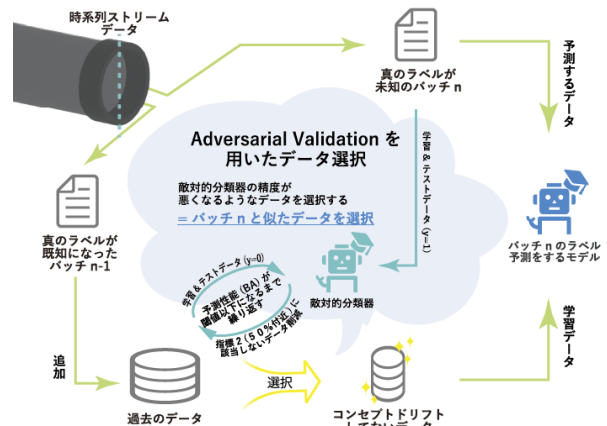


図 1. データ選択手法の流れ

表 1. データセット生成に用いたドリフトパターンの円

	円 2(半径, 座標)	円 3(半径, 座標)	円 4(半径, 座標)
original	0.2, (0.4, 0.5)	0.25, (0.6, 0.5)	0.3, (0.8, 0.5)

表 2. Transition Length が異なる 3 つのデータセットの実験結果(AUC)

	データセット	2,000 件	4,000 件	6,000 件	8,000 件	10,000 件
指標 1	original-tl2000	81.87	84.65	87.49	84.77	82.98
	original-tl6000	73.80	81.77	82.68	83.06	80.21
	original-tl10000	71.43	78.43	79.25	78.31	77.20
指標 2	original-tl2000	72.70	77.41	81.68	83.88	82.98
	original-tl6000	79.46	80.38	83.27	83.01	80.21
	original-tl10000	81.64	84.43	85.58	80.10	77.20

表 3. データ選択手法を用いた実験結果

		ループ終了条件(BA)			全件
		50%	60%	70%	
Original_tl10000	AUC	81.36	77.25	77.20	77.20
	選択件数(10,000)	7026.17	9975.00	10000	10000
	ループ数	10.43	1.067	1.000	-

4. データ選択手法の実験

3. の実験から、継続的にドリフトが起きるデータにおいて大きな性能改善が見込めることが確認できたデータ選択指標 2 を用いて、データ選択手法の効果を確かめる実験を行った。データ選択手法の流れを図 1 に示す。図 1 に示したデータ選択手法では、敵対的分類器の予測性能がループの終了条件である閾値(Balanced Accuracy, 以下 BA と称する)を下回るまで、データ選択指標に従ってドリフトを起きていると思われるデータの少量削減と、敵対的分類器の学習のループを繰り返す。

ここでは、結果の比較を行えるようドリフトパターン original に tl10000 を用いて生成した 3. の実験にも用いたデータセットを利用し、バッチ 2 でバッチ 3 の予測を行う 3. と同様の設定でデータ選択手法の効果を確かめるための実験を行った。データ選択手法には、一度に削減するデータの量を定める閾値 red や、過剰な削減を防ぐための閾値 lim が用意されている。実験には、ループ終了条件の閾値として 50%, 60%, 70% の三つを試し、比較を容易にするため閾値 lim は利用せず、なるべく細やかにデータ選択を行うため閾値 red には 5% を用いた。また、結果にはシードを変えて 30 回実験を行った平均値を利用した。表 3 に示した実験結果から、データ選択手法を利用することで直前の

バッチ全部をモデルの更新に使うよりも、性能のよいモデルを得られることを確認した。

5. まとめと今後の課題

ドリフト環境下において、敵対的分類器を用いて選んだ再学習に適したデータを使ってモデルの更新を行い、データ選択をしない場合よりも性能の良いモデルを得られる可能性があることを示した。本研究の実験には、人工ドリフトデータセットである CIRCLES データセットを用いた。

今後の課題として、データ選択手法の評価を充実させるために他のドリフトデータセットを用いて実験を行いたい。また、敵対的分類器に利用した機械学習モデルの種類を変えて実験を行い、どのような影響があるのかを確かめたい。

参考文献

- [1] J. Pan, V. Pham, M. Dorairaj, H. Chen, and J. Lee. Adversarial validation approach to concept drift problem in automated machine learning systems. CoRR, Vol. abs/2004.03045, , 2020.
- [2] Yuma Konno, Miyuki Nakano and Masato Oguchi. Efficient Data Selection Indicators for Updating Models under Data Drifted Environment. 2022 IEEE International Conference on Big Data (Big Data), 2022.
- [3] Miroslav Kubat and Gerhard Widmer. Adapting to Drift in Continuous Domains. OFAI, Vienna, 1994.