

# コンセプトドリフト対処のための, Adversarial Validation を用いた学習データ選択に関する検討

今野 由麻<sup>†</sup> 中野 美由紀<sup>‡</sup> 小口 正人<sup>†</sup>

<sup>†</sup>お茶の水女子大学 〒112-8610 東京都文京区大塚 2-1-1

<sup>‡</sup>津田塾大学 〒187-8577 東京都小平市津田町 2-1-1

E-mail: <sup>†</sup> {konno.yuma, oguchi}@is.ocha.ac.jp, <sup>‡</sup> miyuki@tsuda.ac.jp

あらまし Adversarial Validation は、機械学習において学習データとテストデータの分布の違いを検出する手法であり、得られたデータ分布の違いを加味して学習データを調整し、より精度の高いモデルを得られるため、最近着目されている。先行研究として、Adversarial Validation をコンセプトドリフトの問題解決のために利用し、時系列データの予測を行う枠組みが発表されている。本論文では、先行研究の手法を拡張し、特徴量選択に加え、時系列の多量なデータから学習に用いるデータ選択を行う方法について検討をする。先行研究で用いられたデータセットを利用した評価実験を行い、特徴量選択を行うよりも精度が向上するケースがあることを確認した。

**キーワード** コンセプトドリフト, Adversarial Validation, 時系列データ

## 1. はじめに

機械学習モデルの利活用が進み、あるタスクを行うモデルが長期的に利用されるシナリオが想定されるようになった。しかし一度学習を行い、良い精度を得られたモデルであっても、使い続けるうちに精度が低下していくことがある。そのような現象の原因として、コンセプトドリフトが知られている。

本研究では、大規模ストリームデータにおいて時間経過などによりコンセプトドリフトが発生する場面で、特定のタスクを行う教師あり機械学習モデルを継続的に入手し、自動的に時系列データの予測を行う枠組みに関する検討を行う。

## 2. コンセプトドリフトとは

コンセプトドリフトとは、動的に変化する非定常な環境において、データの分布が時間経過とともに変化する事象[1]である。ここでは Gama らによる研究[1]を参考にコンセプトドリフトについて簡単に説明する。

コンセプトドリフトの経時変化にはいくつかのパターンがある。例えば、突然次のコンセプトに移り変わるものや、徐々に移行変わるもの、過去のコンセプトが再び起こるものなどである。また、コンセプトドリフト自体にもいくつかの種類があり、リアルドリフトやバーチャルドリフトなどが存在する。今回の提案手法は、データの説明変数の分布が変化するパターンであるバーチャルドリフトへの対処を目的としている。

コンセプトドリフトへ対処するためのアプローチは一般に、1)インスタンス選択、2)インスタンス重み付け、3)アンサンブル学習の3つに分けることができる[2]。そのうちのインスタンス選択の一形態として、case-base editing が存在する。case-base editing の多く

は、ノイズの多いインスタンスや無関係なインスタンス、冗長なインスタンスを削除するものである[3]。

## 3. 先行研究・関連研究

ここでは、Adversarial Validation を利用した2つの研究[4,5]を紹介する。Adversarial Validation とは、学習データとテストデータを見分ける二値分類を行う敵対的分類器を用いて、二つの集団の分布が異なっていることを検出する手法である。この手法は、機械学習コンテストでよく用いられる手法であるが、論文としての発表は数少ない。論文において Adversarial Validation を用いた先行研究を二件示す。

一つ目は、J. Pan ら[4]による Adversarial Validation を用いて特徴量選択を行う方法を含む3通りのコンセプトドリフト適応手法に適用し、比較・検討した、自動でコンセプトドリフトに適応可能なユーザーターゲットシステムの実例である。本論文の提案手法は、この研究の内容を拡張したものとなっている。

二つ目は、Adversarial Validation を用いてバリデーションデータ選択を行った事例[5]である。この例では、Adversarial Validation をバリデーションデータ選択に利用することで、E コマースの購入意図予測タスクにおけるデータの不均衡さに対処している。

## 4. 提案手法

本研究は Pan らによる先行研究[4]を拡張するものである。[4]と同様に、ストリームデータがバッチ単位で処理され、新バッチ処理(バッチ n)が行われるタイミングで直前バッチ(バッチ n-1)のラベルが判明する。

図1に提案手法の概要を示す。本研究の要となるデータの選択を行う方法として、敵対的分類器(GBDT)の

精度(AUC スコア)が与えられた閾値(本手法では 0.8 を採用)より小さくなるまで、分類精度が良いデータの削減と敵対的分類器の再学習を行う。最終的に残ったデータをテストデータのラベルの予測に用いる。この結果、新たに到着したテストデータと分布の近い学習データで学習することでモデルを更新し、コンセプトドリフトの影響を小さくできる。

本提案手法におけるデータ選択手法を行うことの利点として、データ選択時に生成するモデルは再利用されないため、汎化性能を必要とされていないこと、さらにコンセプトドリフトの影響により学習データが母集団全体の代表的な値であるという仮定が成立しなくなるため、ドリフト後のコンセプトを説明可能なデータを選択できることが挙げられる。

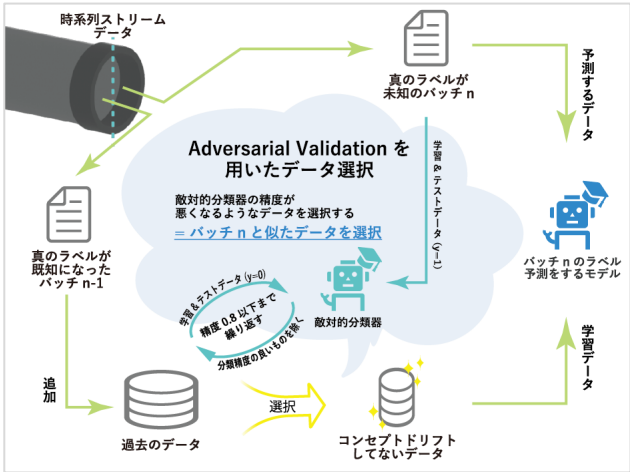


図 1. 提案手法概要

5. 評価実験

5.1. データセット

本研究では、Pan らの先行研究においても利用されていたデータセットのうち、AutoML2 で公開された RL と AutoML3 で公開された B、C、D を用いて実験を行った。これらはいずれもコンセプトドリフトが発生しているデータセットであり、そのサイズについての情報を以下の表 1 に示す。いずれも時系列でバッチとして分割されており、データセット RL はこの中でも比較的軽量のデータセットであることがわかる。

表 1. 利用したデータセットのサイズ

バッチ	1	2	3	4	5
名称	Train1	Test1	Test2	Test3	Test4
サイズ(RL)	31,406	5,000	5,000	14,804	
サイズ(B)	160,055	174,969	160,196	154,034	166,526
サイズ(C)	180,346	193,360	131,574	159,671	183,193
サイズ(D)	153,042	153,124	157,852	153,933	163,830

5.2. データ選択のための閾値

データ選択を行うために、新たに検討しなければならない閾値が二つある。一つ目は、選択するデータの範囲を大まかに決める閾値で、過度なデータ削減の縮小を防ぐ役割も果たす。二つ目は、ちょうど良いデータの削減量を目指して繰り返し Adversarial Validation が行われる中で、一度にどのくらいのデータを削減するのかを決める閾値である。この閾値二つを、以降では閾値 **lim**、閾値 **red** とそれぞれ呼称する。また、データの分類確率としては、足し合わせると 1 になる  $y=0$  に分類される確率と  $y=1$  に分類される確率の二つが得られるが、本研究ではプログラムの簡単のために常に 0.5 以上の数値を利用している。

以下の図 2 は、この閾値を用いてデータ選択が行われる流れを説明したものである。図 2 の 1.では、敵対的分類器の学習用データとテスト用データは SEED を用いてランダムに選択される。2.では、学習したモデルから得た分類確率を使ってデータをソートしている。3.では、青枠で示した部分は閾値 **lim** によって与えられる削減するデータの範囲であり、そのうち特に分類確率の高い閾値 **red** 以上の過去のデータを削除している。分類精度の良いデータを繰り返し少しずつ削減して、敵対的分類器の精度を下げることで、最終的に予測するバッチに近い分布を持つデータを選択することができる。

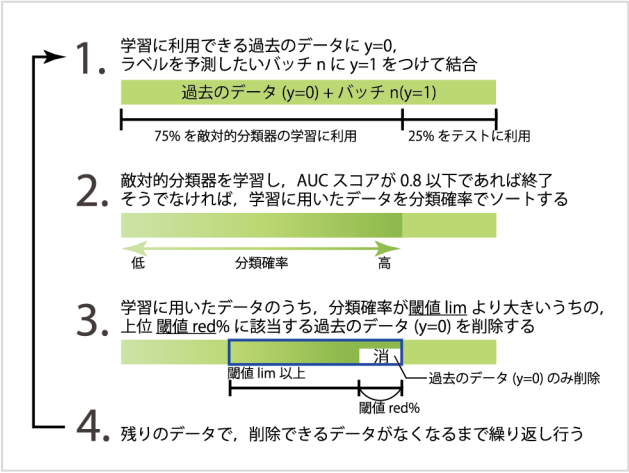


図 2. データ選択の流れ

本来、データの特性により、学習に必要なデータ量の削減とその結果として、必要最小限のコストで学習器の精度向上を図るための閾値は異なると思われる。そこで、今回は Pan らの研究で用いられた 4 つのうちの一つであるデータセット RL を対象として、閾値として適切な値があるか、また閾値を変えた場合の最終的な学習の精度の変化の様子を調べた。

まずは、閾値 `lim` の適切な値を調べるために行った実験の結果を図 3 に示す。このグラフの縦軸は、SEED を変えて 30 回実験して得たバッチ 2,3,4 の予測精度 (AUC スコア) の平均である。この実験では、なるべく細やかなデータ選択を行うために閾値 `red` に関しては 10% を採用し、データ選択を行う対象は予測を行う時点 ( $t=n$ ) で入手できている全てのバッチ (バッチ 1 からバッチ  $n-1$ ) とした。

続いて、閾値 `red` についても同様に実験を行った結果を以下の図 4 に示す。ここでは図 3 の結果で最も良い精度を得られた閾値 `lim`=0.85 を用いて実験を行った。

これらの実験結果から、閾値 `lim` に関しては右肩上がりに精度が向上する傾向があり、小さい値を選びすぎると、データを過剰に削減してしまうことが考えられる。また、閾値 `red` に関しては、右肩下りに精度が低下する傾向があり、大雑把にデータ選択を行うよりも細やかにデータ選択を行った方が精度はよくなる傾向が読み取れる。一方で、2 つの閾値それぞれを変化させたことによる精度の差は小さく、図 3 に示した実験の最大値 68.41% と最小値 68.03% の差は僅か 0.38% であり、図 4 に示した実験の最大値 68.86% と最小値 66.03% の差は 2.83% である。この結果では、データ選択のループによって、この 2 つの閾値による精度の変化はある程度吸収されてしまうことがわかる。

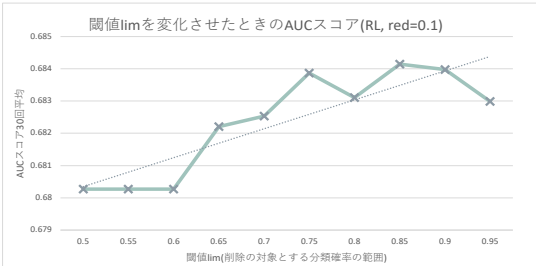


図 3. 閾値 `lim` による精度の変化

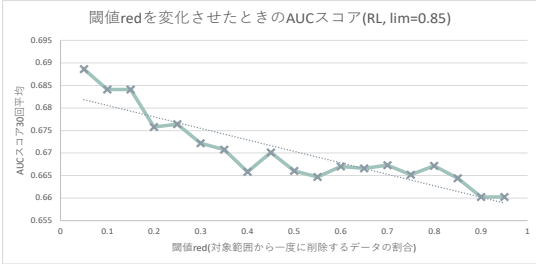


図 4. 閾値 `red` による精度の変化

### 5.3. 取得したログデータについて




各予測においてデータ選択前後のデータのサイズや特徴量の数、データ選択のループの回数を取得した。また、各ループにおいて敵対的分類器の精度や削減されるデータの件数を取得した。

### 5.4. データ選択を行う過去のバッチの選び方

ここでは、データセットの特性とデータ選択を行う過去のバッチの選択方法の関係を観察するために、データセット RL を例にして、3 通りの方法を検討した。また、提案手法であるデータ選択のほかに、比較対象としてデータ選択を行わずにそのまま学習を行なったモデルであるベースラインを用意した。この実験で用いた閾値は、5 章 2 節の実験で最も良い精度を得られた `lim`=0.85, `red`=0.05 である。

次に、検討したバッチの選び方について、以下の表 2 に示す。選び方 (a) の場合は、精度の向上は期待できるが膨大なデータを扱うため学習コストが高くなる。また、発生したドリフト次第では、古いデータを加えることは精度低下の要因となる可能性がある。選び方 (b) の場合は、新しいコンセプトにも対応できる可能性があるが、古いコンセプトの再発への対応力は (a) に比べ低くなると考えられる。選び方 (c) の場合は、新しいコンセプトに対応できない。今回の実験で利用するデータセット RL の場合は、選び方 (a), (c), (b) の順番で学習に利用できるデータのサイズが大きくなる。

表 2. データ選択を行うバッチの選び方

名称	説明	バッチ n の予測に用いる過去のデータ
(a)	入手可能な全データを利用	 バッチ 1 ... バッチ n-1
(b)	直前のバッチのみ利用	 バッチ n-1
(c)	一番古いバッチのみ利用	 バッチ 1

続いて、実験結果を示した表 3,4,5,6 と図 5 の紹介と、その結果の考察を行う。表 3 は、データ選択とベースラインそれぞれにおいて、SEED を変えて 30 回実験して得たバッチ 2,3,4 の予測精度 (AUC スコア) の平均であり、表 4,5,6 はデータ選択におけるバッチ 2,3,4 それぞれの予測に関する表である。表 4,5,6 には、各バッチの予測精度 (AUC スコア) やデータ選択前後のデータ数、繰り返し行われるデータ選択のループが何周で止まったのか (ループの回数) などが示されている。表 4 の全ての選び方で結果が共通な理由は、学習に利用するバッチがこの時点では全て同じであるためである。図 5 は、表 4,5,6 にも記載されている各バッチの精度に、ベースラインの精度を加えたグラフである。

表 3 の結果から、データ選択を行う場合のバッチの選び方 3 通りの精度を比較してみると、選び方(a), 選び方(c), 選び方(b)の順に精度が良く、学習に利用できるデータの量が多いときほど精度が上がる傾向にある。また、図 5 の選び方(b)の精度が安定していることから、この実験で利用したデータセット RL は比較的安定している、もしくは安定して推移するデータセットである可能性が推測される。バッチ 4 の予測では、選び方(c)の精度が良いことから、過去のコンセプトが再発している可能性があり、古いデータも将来の予測に有用な可能性があることがわかる。よって、データセットの特性や最終的な学習のタスクで対処したいドリフトのタイプによって学習に用いるバッチの選び方を変えるべきであると考えられる。

次に、本実験で観測された提案手法の課題点について考察する。表 4 や図 5 の結果から、データ選択とベースラインを比較してみると、選び方(b)と選び方(c)でデータ選択を行うことで精度が改善したことがわかる。しかし、選び方(a)のバッチ 3 とバッチ 4 の予測ではベースラインと比較して精度が下がってしまっている。その点について注目して、表 4,5,6 を確認すると、選び方(a)のバッチ 3 とバッチ 4 の予測で他よりも多くループが回り、全体のうちの少しのデータのみが学習で利用されている。この場合、過剰なデータ選択に陥り、本来学習に役立つデータまで取り除いてしまったことが考えられる。予測精度を下げる可能性のあるデータだけを適切に取り除いて学習を行うために、適切な閾値の調整方法や、過剰なデータ選択が行われないためのアルゴリズムの改善を検討していく必要がある。

表 3. 各バッチの精度の平均

	バッチの選び方	精度(AUC スコア)
データ選択	(a)	68.86
	(b)	<b>64.83</b>
	(c)	<b>65.49</b>
ベースライン (データ選択なし)	(a)	<b>69.84</b>
	(b)	64.35
	(c)	64.77

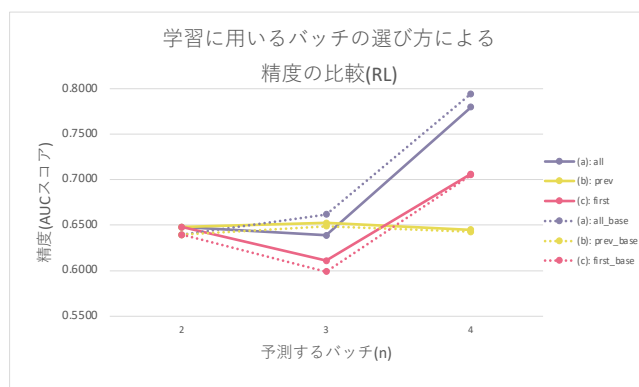


図 5. 学習に用いるバッチの選び方による精度の比較(RL, 実線: データ選択, 点線: データ選択なしのベースライン)

表 4. データ選択を行うバッチの選び方を変えた場合のデータセット RL のバッチ 2 の予測結果

バッチの選び方	精度(AUC スコア)	選択前のデータ数	選択後のデータ数	選択された割合	ループの回数
(a), (b), (c)	64.79	31,406	16,324.00	51.98%	15.23

表 5. データ選択を行うバッチの選び方を変えた場合のデータセット RL のバッチ 3 の予測結果

バッチの選び方	精度(AUC スコア)	選択前のデータ数	選択後のデータ数	選択された割合	ループの回数
(a)	63.86	36,406	4,312.13	11.84%	65.40
(b)	65.23	5,000	4,191.53	83.83%	26.27
(c)	61.08	31,406	14,892.03	47.42%	17.20

表 6. データ選択を行うバッチの選び方を変えた場合のデータセット RL のバッチ 4 の予測結果

バッチの選び方	精度(AUC スコア)	選択前のデータ数	選択後のデータ数	選択された割合	ループの回数
(a)	77.94	41,406	10,152.27	24.52%	48.23
(b)	64.47	5,000	4,998.97	99.98%	1.67
(c)	70.61	31,406	29,015.67	92.39%	2.53

表 7. 他手法との比較を行った実験結果(各バッチの平均)

データセット	データ選択	特徴量選択	ベースライン (データ選択等なし)
データセット RL	64.83	52.83	64.35
データセット B	56.67	60.07	57.06
データセット C	67.81	69.25	67.40
データセット D	62.69	62.13	63.17

表 8. 他手法との比較を行った実験結果詳細(データセット RL, データ選択)

予測するバッチ	精度(AUC スコア)	選択前のデータ数	選択後のデータ数	ループの回数
n = 2	64.79	31,406	16,324.00	15.23
n = 3	65.23	5,000	4,191.53	26.27
n = 4	64.47	5,000	4,998.97	1.67

表 9. 他手法との比較を行った実験結果詳細(データセット RL, 特徴量選択)

予測するバッチ	精度(AUC スコア)	選択前の特徴量数	選択後の特徴量数	ループの回数
n = 2	52.36	22	6.97	9.07
n = 3	57.41	22	15.00	4.00
n = 4	48.73	22	14.53	4.23

表 10. 他手法との比較を行った実験結果詳細(データセット B, データ選択)

予測するバッチ	精度(AUC スコア)	選択前のデータ数	選択後のデータ数	ループの回数
n = 2	56.20	160,055	151,698.27	28.07
n = 3	57.28	174,969	115,128.07	29.87
n = 4	56.63	160,196	139,483.27	21.00
n = 5	56.56	154,034	150,245.43	22.83

表 11. 他手法との比較を行った実験結果詳細(データセット B, 特徴量選択)

予測するバッチ	精度(AUC スコア)	選択前の特徴量数	選択後の特徴量数	ループの回数
n = 2	59.94	24	21.00	2.00
n = 3	60.13	24	20.90	2.03
n = 4	60.06	24	21.00	2.00
n = 5	60.13	24	21.00	2.00

表 12. 他手法との比較を行った実験結果詳細(データセット C, データ選択)

予測するバッチ	精度(AUC スコア)	選択前のデータ数	選択後のデータ数	ループの回数
n = 2	68.97	180,346	173,616.83	47.33
n = 3	68.89	193,360	176,285.27	75.50
n = 4	70.66	131,574	131,573.50	1.17
n = 5	62.71	159,671	159,648.47	2.40

表 13. 他手法との比較を行った実験結果詳細(データセット C, 特徴量選択)

予測するバッチ	精度(AUC スコア)	選択前の特徴量数	選択後の特徴量数	ループの回数
n = 2	62.19	64	51.00	3.00
n = 3	73.16	64	52.20	2.80
n = 4	72.70	64	51.00	3.00
n = 5	68.94	64	54.60	2.40

表 14. 他手法との比較を行った実験結果詳細(データセット D, データ選択)

予測するバッチ	精度(AUC スコア)	選択前のデータ数	選択後のデータ数	ループの回数
n = 2	61.56	153,042	69,774.87	29.17
n = 3	63.28	153,124	70,273.40	31.27
n = 4	62.23	157,852	70,249.20	33.20
n = 5	63.70	153,933	71,622.37	33.17

表 15. 他手法との比較を行った実験結果詳細(データセット D, 特徴量選択)

予測するバッチ	精度(AUC スコア)	選択前の特徴量数	選択後の特徴量数	ループの回数
n = 2	62.14	72	64.00	2.00
n = 3	62.89	72	64.00	2.00
n = 4	61.36	72	64.00	2.00
n = 5	62.13	72	64.00	2.00

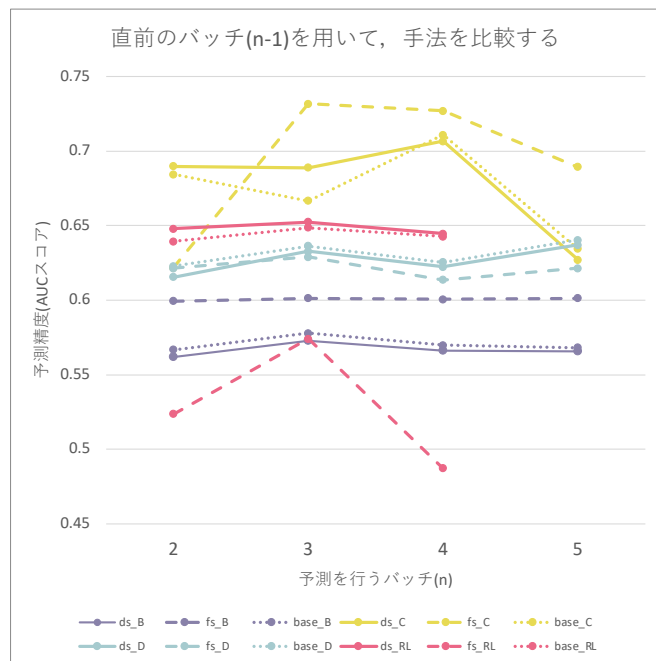


図 6. 他手法との比較を行った実験結果(実線: データ選択(ds), 波線: 特徴量選択(fs), 点線: データ選択等なしのベースライン(base))

### 5.5. 他手法との比較を行った実験結果

提案手法であるデータ選択以外の手法として, Panらによる先行研究のうち, 敵対的分類器に GBDT を利用したものを再現した特徴量選択と, 学習データに手を加えていないベースラインを用意して実験を行った. 実験にはデータセット RLに加えて, データセット B, C, D も用いた. これまでの実験と同様に, 今回の実験も SEED を変えて 30 回実験して得た結果を平均している. また, 本節の実験では, 全ての方法に共通してバッチ  $n$  の予測のためにバッチ  $n-1$  のみ(バッチの選び方(b))を利用した.

実験結果を表 7 から表 15 と図 6 に示す. 表 7 には, 予測を行った各バッチの精度(AUC スコア)の平均をデータセットごとに記載した. 表 8 から表 15 は, それぞれのデータセットに関して, データ選択と特徴量選択の各予測の結果について示したものである. これらの表には, 各バッチの予測精度(AUC スコア)やデータ選択前後のデータもしくは特徴量数, 繰り返し行われるデータ選択のループが何周で止まったのか(ループの回数)が記載されている. 図 6 には, 表 8 から表 15 にも記載されているデータ選択と特徴量選択の各バッチの精度に加えて, ベースラインの各バッチの精度をグ



ラフに示した。

表 8 を例にして、実験結果の読み取り方を説明する。まずループの回数に注目してみると、 $n=4$  の時に早期にループが終了しており、データ選択前後でデータの件数がほとんど減っていないことがわかる。このようにデータ選択がほとんど行われていない場合には、ドリフトが発生していない可能性や、閾値が適切でないなどの要因で提案手法のアルゴリズムが効果的に動作しなかった可能性などが考えられる。

本節の以降では、各データセットについて実験結果を評価する。

#### 5.5.1. データセット RL について

表 8 および表 9 と図 6 から、データ選択を行うことでベースラインより僅かに精度が向上していることがわかるが、ベースラインからの改善幅は余り大きくない。また、特徴量選択を行なった場合は精度が大幅に下がることがわかる。ベースラインの精度が安定していることなどから考えて、このデータセットはバッチ間で急激なドリフトが生じるようなデータセットではなく、比較的穏やかなドリフトが起きていると推測される。

#### 5.5.2. データセット B について

表 10 および表 11 と図 6 から、データ選択では過剰なデータ削減によって精度が低下している一方で、特徴量選択では安定して高い精度を出すことができていくことがわかる。このデータセットで発生しているドリフトは特徴量に強く関連していることが推測される。

#### 5.5.3. データセット C について

表 12 および表 13 と図 6 から、どの手法に関しても精度がバッチごとに上下しており、ドリフトが起きていることが推測できる。データ選択では、バッチ 2 とバッチ 3 の予測において、バッチ 4 とバッチ 5 の予測より明らかにデータ選択のループが多く回っており、実際にその時の精度がベースラインを上回っていることが確認できた。

#### 5.5.4. データセット D について

表 14 および表 15 と図 6 から、データ選択では、大体半分くらいのデータが選択されているが、ベースラインと比較するとどのバッチでも過剰な削減が行われていることがわかる。特徴量選択に関しても同様の結果となっている。

提案手法は特徴量選択やベースラインにおける結果よりも一部のデータセットで良い結果を出すことができていく。しかし、ベースラインからの精度向上の幅は非常に小さく、この結果の原因究明のための追加実験や、各データセットに対して適切な閾値を求める方法について検討する必要がある。特にデータセット

B, C, D のデータ選択に関しては、データを過剰に削減して精度を低下させているバッチがある。まずはこの過剰な削減が、提案手法の範疇である閾値の調整によってどのくらい改善されるのかを調査し、必要に応じてデータ選択のアルゴリズム自体を改良する必要がある。

## 6. まとめと今後の課題

本研究では、自動でコンセプトドリフトに適応可能なシステムに関する先行研究の手法をベースとして、Adversarial Validation をデータ選択手法に適応する手法を新たに提案した。

評価実験から大きな改善は得られなかったが、これはデータセットの性質など、コンセプトドリフトがどのように起きているかが一つの要因として考えられる。

今後の課題として、学習データに用いるバッチの選択方法やデータセットごとに適切な閾値が変わる可能性が本研究から示唆されたため、適切な閾値の検討を行うための更なる実験が必要である。その際に、過剰なデータ選択によってかえって精度を低下させてしまう課題がどの程度改善するかを観察し、必要に応じてデータ選択のアルゴリズムを改良していきたい。

また、Pan らによる先行研究と同様に GBDT 以外の DT などの機械学習モデルも敵対的分類器として利用できるよう実験と検討を行いたいと考えている。

## 参 考 文 献

- [1] Gama, João et al, “A survey on concept drift adaptation”, ACM Computing Surveys (CSUR) 46 (2014): 1 - 37.
- [2] Ning Lu et al, “A concept drift-tolerant case-base editing technique”, Artificial Intelligence 230 (2016) 10.
- [3] Cheng-Jung Tsai et al, “Mining decision rules on data streams in the presence of concept drifts”, Expert Systems with Applications 36 (2009): 1164 - 1178
- [4] Jing Pan et al, “Adversarial validation approach to concept drift problem in automated machine learning systems”, CoRR, Vol. abs/2004.03045, , 2020.
- [5] Shotaro Ishihara et al, “Adversarial Validation to Select Validation Data for Evaluating performance in E-commerce Purchase Intent Prediction”, <https://sigir-ecom.github.io/ecom21DCPapers/paper3.pdf>, SIGIR eCOM'21

## 謝 辞

本研究は一部を JST CREST JPMJCR1503 の助成、一部を JSPS 科研費 18K11318 の助成を受けたものです。