

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考察

今野由麻[†] 中野美由紀[‡] 小口正人[†]

お茶の水女子大学[†] 津田塾大学[‡]

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。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である [1]。

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考察を行う。

2. 先行研究・関連研究

ここでは、Adversarial Validation を利用した 2 つの研究 [2,3] 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。

J. Pan ら [2] は、Adversarial Validation を用いて特徴量選択を行う方法を 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ゲーティングシステムの事例となる。

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 [3] である。

3. 提案手法

本研究は、Pan らによる先行研究 [2] を拡張するものである。[2] 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 $n-1$) の真のラベルが判明するという前提を用いる。

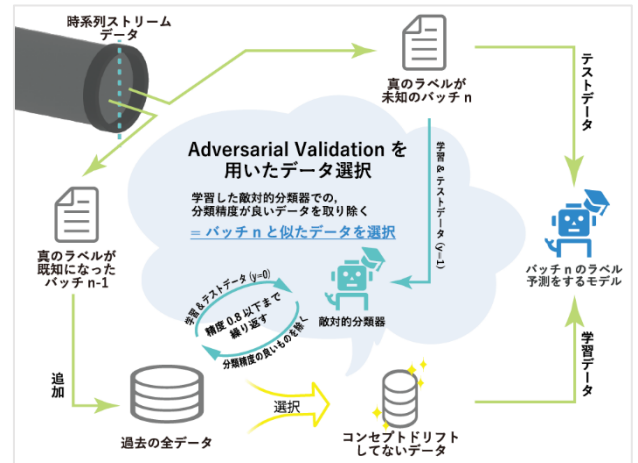


図 1. 提案手法概要

図 1 に提案手法の概要を示す。提案手法の要となるデータの選択手法として、敵対的分類器の精度 (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除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。

4. 評価実験

4.1 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 RL を用いて実験を行った。このデータセットは、AutoML2 という機械学習コンテストのために公開されたデータであり、あらかじめコンセプトドリフトが発生していることがわかっている。このデータセットの情報を以下の表 1 に示す。

表 1. データセット RL

バッチ	1	2	3	4
名称	Train	Test1	Test2	Test3
サイズ	31,406	5,000	5,000	14,804

4.2 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、過度なデータサイズ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除量を目指して繰り返し Adversarial Validation が行われる中で、一度でどのくらいのデータを削除するのかが決める閾値である。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。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。

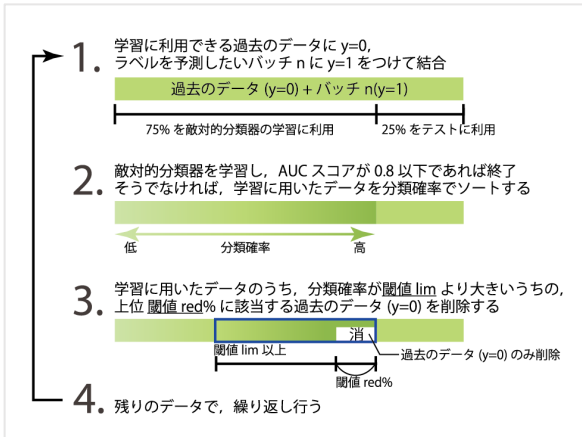


図 2. データ選択の流れ

閾値 *red* に関しては、なるべく細やかなデータ選択を行うために、以降の実験では 10% を利用した。また閾値 *lim* の適切な値を調べるために、データセット RL を用いた実験を行った結果を図 3 に示す。この実験結果においては、*lim*=0.85% が最も良い精度を得られたが、他のデータセットを用いてさらなる検討を行う必要がある。またこの実験結果から、閾値 *lim* による精度の差は、データ選択のループによってある程度吸収されてしまうことがわかる。

4.3 他手法との比較を行った実験結果

バッチ 2,3,4 の予測精度(AUC スコア)の平均を、各手法で SEED を変えて 30 回実験して得た結果を表 2 に示す。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た。また、敵対的分類器を利用する手法では、共通して GBDT を利用した。

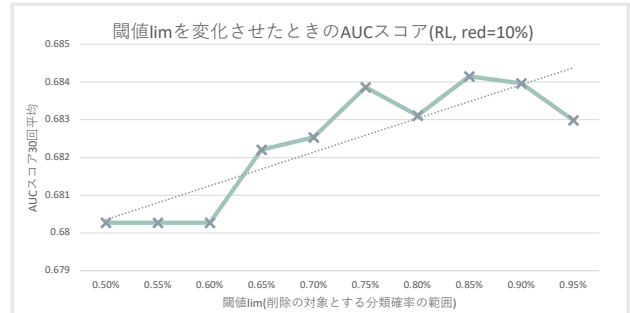


図 3. 閾値 *lim* を変化させた場合の AUC スコア

提案手法がベースラインと比較してあまり変わらない結果となった理由として、利用したデータセット RL の特性が影響していると考えられる。

表 2. AUC スコア 30 回平均

データ選択	特徴量選択	ベースライン
68.41	53.02	69.84

5 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などが要因として考えられる。

今後の課題として、実験に利用したパラメータの改善および本手法の効果測定のための、異なる特性を持ったデータセットを用いた実験が必要である。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アルゴリズムを検討したいと考えている。

参考文献

- [1] Gama, João et al, “A survey on concept drift adaptation”, ACM Computing Surveys (CSUR) 46 (2014): 1 - 37.
- [2] Jing Pan et al, “Adversarial validation approach to concept drift problem in automated machine learning systems”, CoRR, Vol. abs/2004.03045, , 2020.
- [3] Shotaro Ishihara et al, “Adversarial Validation to Select Validation Data for Evaluating performance in E-commerce Purchase Intent Prediction”, <https://sigirecom.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。