小口研究室　研究紹介　(2022年度)
(お茶の水女子大学理学部情報科学科)

敵対的分類器を用いた再学習用データ選択手法の研究
（研究担当:今野　由麻）

研究背景

- 大量のデータが生成され、機械学習が広く利用されるようになった
- 一度学習したモデルでも、使い続けると精度が低下することが知られている
  → このような現象の原因として、コンセプトドリフトがある
- 精度が低下するたび、新しいデータでモデルを学習し直す等の簡単な手法(図1)
  → 新しいデータが学習に最適とは限らない & 過去の有用なデータを捨ててしまう

大量のストリームデータが生成される場面において、<u>学習データを適切に選んだモデル</u>を継続的に入手する方法を考えたい

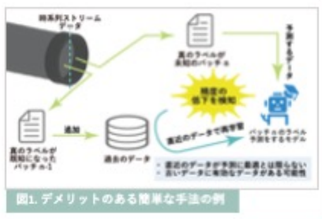


図1. デメリットのある簡単な手法の例

提案手法

- 本手法は、Panらによる先行研究[1]を拡張したものである
- 複数の時系列データのバッチを持ち、真のラベルが未知のバッチnが入手された時点でバッチn-1の真のラベルが得られるような状況を想定している
- 真のラベルが未知のバッチnの真のラベルを予測する機械学習モデルを得ることを目的のタスクとしている

- 提案手法の流れ(図2)
  1. バッチnの真のラベルの予測を行うために利用できるデータとして、バッチ1からバッチn-1が過去のデータに保管されている
  2. 真のラベルが未知のバッチnに敵対的分類器のための新しいラベルy=1を付与し、過去のデータに対してy=0を付与する
  3. 真のラベルが未知のバッチnと過去のデータを見分けるような二値分類器である敵対的分類器を学習する
  4. 敵対的分類器の精度(BA : Balanced Accuracy)が閾値以下になるまで、過去のデータからドリフトしたデータを少量省くことと敵対的分類器の再学習を繰り返す
     → 敵対的分類器の精度が悪くなるようなデータを選択する(選択指標2)ことで、予測したいバッチと似たデータを抽出することができる
  5. 過去のデータのうち省かれずに残ったものを学習データとして、真のラベルが未知のバッチnのラベルを予測するモデルを学習する

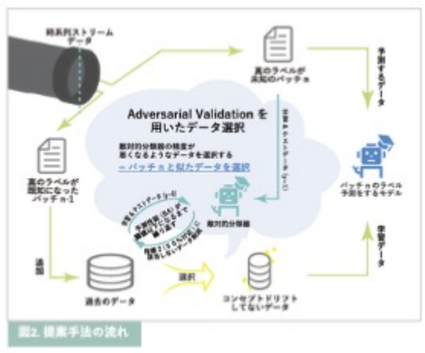


図2. 提案手法の流れ

[1] Jing Pan et al, "Adversarial validation approach to concept drift problem in automated machine learning systems", CoRR, Vol. abs/2004.03045, , 2020.

CIRCLESデータセット

- CIRCLESデータセット[2]はKubatらによって提案された人工ドリフトデータセット
  - 特徴量は0-1の範囲の二次元空間上の座標で、クラスは正もしくは負のどちらか
  - あるカテゴリのデータが移動(ドリフト)する
  - ここでは常にドリフトが進行する設定を利用
  - 複数個の円で構成されたパターンを使用(図3)

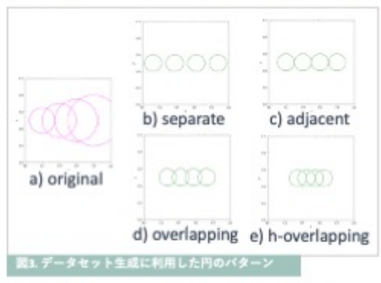
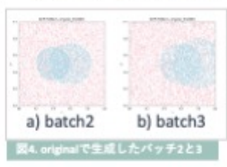


図3. データセット生成に利用した円のパターン

- 図3のうちのoriginalを利用して生成したデータセット(バッチ2と3)の様子(図4)
  - バッチ2はパターンの左から2つ目と3つ目の円、バッチ3は3つ目と4つ目の円で生成されている
  - 各バッチは10,000件のデータで構成



特徴量である座標上に正のクラスをピンク、負のクラスを水色でプロット
図4. originalで生成したバッチ2と3

実験結果

- CIRCLESデータセットを用いてデータ選択手法の効果を確認する実験を行った
  - 5つの円のパターンを用いてそれぞれ生成したデータセットのバッチ2を使って、バッチ3の予測を行った
  - 実験結果を示した表1では、データ選択手法(図2)のループの終了条件として50%/60%/70%の3つを試した結果と、比較のためにデータ選択を行わずにバッチ2全件でバッチ3の予測を行った結果を掲載
  - 実験結果は全て、シードを変えて30回実験を行った結果の平均値を記載

表1. データ選択手法を評価する実験結果

		ループ終了条件(BA) 50%	60%	70%	全件
original	AUC	**81.36**	77.25	77.20	77.20
	選択件数	7026.17	9975.00	10000	10000
separate	AUC	**84.12**	79.55	74.28	74.28
	選択件数	7218.50	8943.53	10000	10000
adjacent	AUC	**85.90**	80.88	75.79	75.79
	選択件数	7236.67	8818.53	10000	10000
overlapping	AUC	**86.97**	85.92	78.89	78.89
	選択件数	7032.47	8783.73	10000	10000
h-overlapping	AUC	87.74	**88.01**	85.41	85.41
	選択件数	6858.40	8533.93	10000	10000

  - 全ての場合で、データ選択を行わない(全件)ときよりもデータ選択を行うことで性能が改善
  - データの変化が大きいとき(=パターンの円の変化大)のほうが、データ選択による性能改善の幅が大きい傾向
    ex) separate, adjacent
  - データごとに最適なデータ選択件数が異なることが予想されるが、今回の結果の多くで7,000件前後が選択

まとめと今後の課題

- 先行研究の手法をベースとして、Adversarial Validationを用いて再学習用データ選択を行うことで自動でドリフトに適応可能なシステムを提案
  →人工ドリフトデータセットであるCIRCLESを用いて手法の評価を行った
- 他のデータセットを用いてデータ選択手法の評価をさらに行いたい
- 敵対的分類器のモデルを変えて、提案手法の有効性を確認したい

[2] M. Kubat, and G. Widmer, "Adapting to Drift in Continuous Domains", OFAI, Vienna (TR-94-27), 1994

サーバシステムの性能データ収集および転送効率化に向けた改善案の評価
（研究担当:飯山　知香）

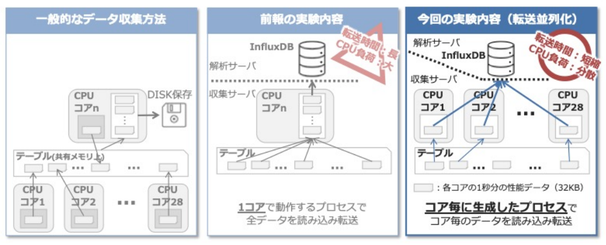
研究概要

<u>背景</u>
- クラウド環境をはじめとした、多数台サーバの共有利用や分散処理利用に関する需要の増加
  - サーバの負荷分散や、システムやアプリケーションのチューニングを行うには、各サーバの低レイヤを含めた性能データを低オーバヘッドで収集してリアルタイムに分析・提示する手法が必要
- 多数台のサーバのデータを一元的にリアルタイムで分析する際、分析対象のサーバとその分析を行うサーバは分割されることが多い
  - Linuxの性能データなどの時系列データはデータサイズが比較的大きく、扱う際のオーバヘッドが大きくなる可能性がある

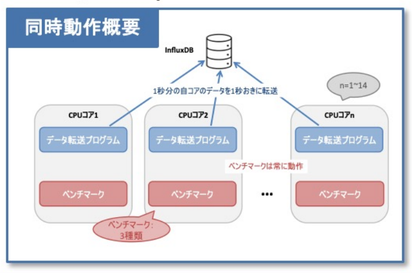
<u>目的</u>
✓ 効率的に性能データを収集・転送する手法の実現

実験概要

<u>転送処理並列化</u>
✓ 前報の実験：複数コアの情報を1つのコアで転送する方式では、コア数が増えると転送処理が追い付かなくなる
＝転送の効率化が必要
✓ 今回の実験：各コアの情報を自コアで転送する並列化を実行し、並列化の効果を検証

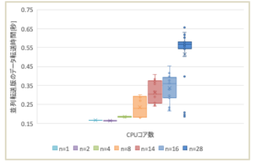


<u>ベンチマーク同時動作</u>
✓ 同一CPUコア上で動作する転送処理とベンチマークの相互影響を調査
✓ ベンチマーク(CPU負荷)：UnixBench(dhry2reg、whetstone-double)、7-Zip



実験結果

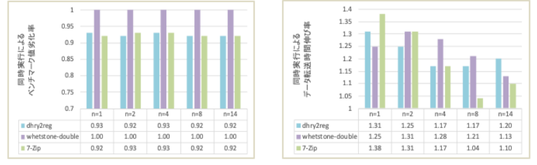
<u>転送処理並列化</u>
- データ転送時間の比較(n=1~28)
  - ✓ 転送時間の短縮（＝並列化による効果）
  - ✓ 1/nに短縮はできていない（＝並列化のオーバヘッド）
  - ✓ コア間で転送時間にばらつき（＝InfluxDB側のCPU負荷率やI/O負荷がネックになっている可能性）

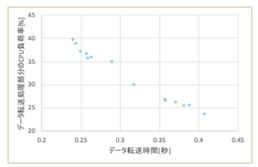


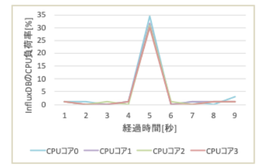
CPUコア数	n=1	n=2	n=4	n=8	n=14	n=16	n=28
平均[秒]	0.17	0.17	0.19	0.24	0.31	0.34	0.51
1コア転送版の約1/○	1/1	1/2	1/3	1/5	1/7	1/7	1/8

- データ転送処理部分のCPU負荷率(n=14)
  - ✓ 転送時間が長くなるほどCPU負荷率が低い（＝Idle時間が増加）
- InfluxDBのCPU負荷率・I/O負荷(n=14)
  - ✓ 処理中はCPU負荷100%でCPUネック（＝転送時間のばらつきの要因）
    - ✓ 4コア全てが同時に30%程度
    - ✓ 転送時間は0.32秒
  - ✓ I/Oに余裕があることはfioベンチマークで確認済み

<u>ベンチマーク同時動作</u>
- ベンチマーク値への影響
  - ✓ ベンチマークごとに異なる
  - ✓ CPUコア数には依存しない
- データ転送時間への影響
  - ✓ ベンチマークによらず伸びた
  - ✓ CPUコア数の増加に伴い伸び率減少（※本実験環境に依存）







今後の課題

- InfluxDBサーバのCPUコア数増加
- 転送プログラムの改善
- ベンチマークによる相互影響の度合いの詳細分析
- CPUコア数増加によるデータ転送時間の伸び率変化の詳細分析

モバイル端末からのイベント情報検索における Geo-indistinguishability を用いたユーザ位置匿名化の評価
（研究担当:石神　京佳）

研究背景

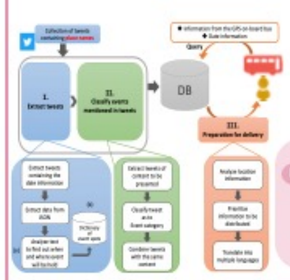
✓ SNS上には固定的なメディアに載っていない大小様々なイベント情報が存在
  - SNSの利用目的「2位：情報収集目的」
✓ データの収集・分析技術の飛躍的な進歩
  - 様々なソースから膨大なデータを処理及び分析することが可能
  - ユーザデータを利用した情報推薦サービスへの応用

目的
SNS上のデータを利用した安心安全な推薦システム
ユーザデータを利用した推薦精度の向上

先行研究と研究方針

SNS データを用いた場所と時間を考慮するイベント検索手法の提案と評価
(お茶大・工藤ら[DICOMO2018])

✓ 時間移動するユーザにリアルタイムにSNS上の有益な情報を配信



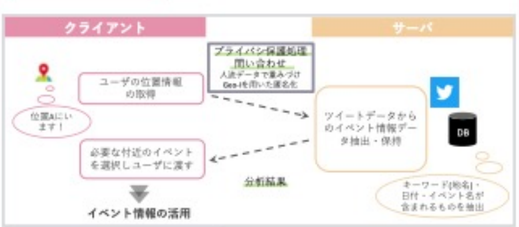
1. Twitter上の地名に紐づいたツイートを収集
   →イベント情報の抽出
2. イベント情報を分析しイベントを分類
3. ユーザの位置情報をもとに有益度順に提示

ユーザのプライバシを脅かす可能性
Ex プライベートなツイートデータ、位置情報
- 情報保有者（サーバ）が一方的にユーザデータを利用
- 個人情報を分析に利用される際のプライバシリスク

✓ **サーバに渡すユーザデータへのプライバシ保護処理が必要**
✓ **様々なプライバシ保護のためのデータ処理の研究が行われている**

Ex 差分プライバシ、匿名化、暗号化

提案システム

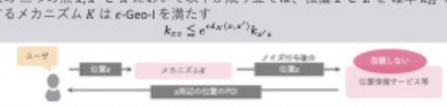


Geo-Iを用いた位置匿名化手法

Geo-I(Geo-indistinguishability)について
- 大まかな位置は知られても良いが、具体的な位置は知られないという要件を適切に表した指標
- Xをユーザの位置の集合、ZをメカニズムKによって曖昧化した結果得られる位置の集合とする

任意の二つの点$x, x' \in X$において以下が成り立てば、位置$x \in Z$を確率k_{xz}で出力するメカニズムKはϵ-Geo-Iを満たす

$$k_{xz} \leq e^{\epsilon d_X(x,x')} k_{x'z}$$



本研究ではBordenabe *1らが提案したユーザ位置が$x \in X$であるπ_xを用いたメカニズムKを使用

人流データを用いた重みづけをπ_xとし
ユーザ位置の匿名化を保護度合いϵで行う

*1「Nicolas E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal geo-indistinguishable mech- anisms for location privacy, 2014.」

実験

✓ **求めたϵ-Geo-Iを満たす確率分布に従いダミー位置を選択、問い合わせの位置として使用**
- サーバにはダミーとして選択されたメッシュ中心位置の緯度経度のみ渡される
- 新宿駅周辺と新橋駅周辺における 500m 四方の分割地域メッシュ 20 区画の人流データを用いて重み付けを行ない、ダミー位置として各区画が選ばれる確率を調査

ϵの値を調整することで・・・
「より人のいる場所／ユーザ位置から近い場所を問い合わせの位置として選んで欲しい」
というユーザの要件に柔軟に対応することが可能に

今後の課題

✓広い範囲、時間帯別や年齢別の人流データを使用
　人流データ「モバイル空間統計」の利用、連続する広い範囲に本手法を適応させる
✓過去のイベントの発生情報を利用しリアルなイベント開催地分布に応じた場合分け
✓ユーザへの情報推薦に位置情報以外のユーザデータの使用を検討

さらにユーザの要望に合うようなイベント情報の推薦を目指す

https://www.is.ocha.ac.jp/~oguchi_lab/　　oguchi@is.ocha.ac.jp