小口研究室　研究紹介　(2025年度)
(お茶の水女子大学理学部情報科学科)

分散データ連携基盤上の改ざん特定・修復可能な手法に関する研究
（研究担当:堀　遥）

研究背景

脱炭素社会実現に向け，サプライチェーン全体でのカーボンフットプリント(CFP)管理基盤の需要増加

カーボンフットプリント(CFP：Carbon Footprint of Product)とは… 製品のライフサイクルの各過程で排出された温室効果ガスを製品単位で表示する仕組み

- サプライチェーン上の企業を分散システムで連携 → CFPを算出・管理
- CFPデータは環境規制への対応などで改ざんリスクがあり，信頼性の担保が必要

課題
- サプライチェーンという複雑な環境下における改ざんは他製品にも影響を及ぼす
  → 基盤の信頼性と可用性維持のためには，**改ざんの特定と修復** が不可欠

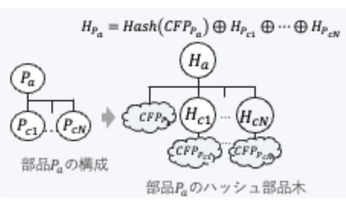
方針
① **ハッシュ部品木によるデータ検証で改ざん特定**
② **特定した改ざんをWhat-If分析フレームワークを用いて効率的に修復**

提案手法　**ハッシュ部品木による改ざん特定**

CFPデータに発生した改ざんを特定可能にする
ハッシュ部品木をブロックチェーンに保管

- 定義：部品の構造に従いCFPのハッシュ値をXORで統合

$$H_{P_a} = Hash(CFP_{P_a}) \oplus H_{P_{c1}} \oplus \cdots \oplus H_{P_{cN}}$$



部品P_aの構成　　部品P_aのハッシュ部品木

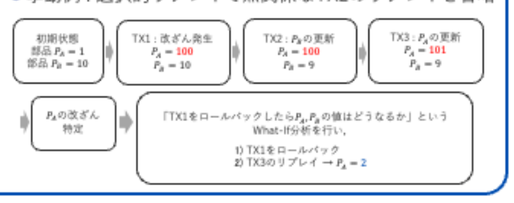
- 改ざん特定手法：改ざんの影響をXORの自己反転性 $(a \oplus a = 0)$で除去して改ざん判定

What-If分析による改ざん修復

What-If分析とは…
データ記録や取引履歴における仮想的な変更をシミュレートし，変更された条件下での結果を予測

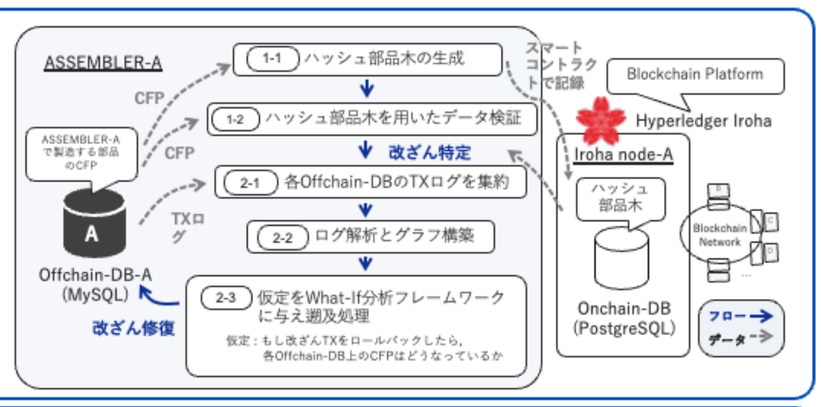
「もしCFPの値に改ざんが発生していなかったら，現在のデータはどうなっているか」という仮定を分析

- 挙動例：選択的リプレイで無関係なTX2のリプレイを省略



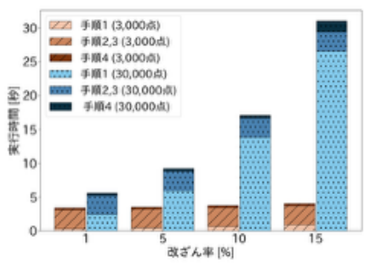
提案システムの全体像

- ローカルDBであるOffchain-DB上のCFPデータに発生した改ざんを特定・修復
- ハッシュ部品木には，Blockchainの恩恵で高信頼なOnchain-DBに保管
- 各ASSEMBLERはBlockchain Networkで接続



評価　**時間的コストの調査**

- 評価指標：
  CFP更新TX数(10,000件)，部品の総数(3,000/30,000点)，重複率(10%)，改ざん率(1/5/10/15%)
- 改ざん特定：**特定率100%を確認**
  総数・改ざん率に対して増加傾向 ➡ 処理対象の部品数への比例を確認
- 改ざん修復：**正確な修復を確認**
  同様に増加傾向であるが安定して約3秒
  ➡ 高いスケーラビリティにより，What-If分析がサプライチェーン上のデータ連携基盤におけるデータ修復に有効であると示唆された



結論

CFP管理シナリオは，データ更新や整合性確認の頻度が相対的に低く，リアルタイム性よりもデータの正確性や信頼性が重要
➡ 提案する改ざん特定・修復手法は，CFP管理基盤実運用において十分に許容範囲内な性能

プライバシ保護を考慮した機械学習モデルによる異常検知に関する研究
（研究担当:森　仁美）

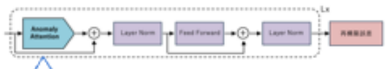
研究背景

- プライバシ保護されたデータに対して、深層学習による異常検知を行う技術が要求
- 時系列データに含まれる異常は5種
  →検知難易度が異なり、特定の異常が多く検知されている恐れ
- プライバシ保護の観点を考慮した、異常ごとの性能評価は不十分

- 各異常の特徴がAnomaly Transformerでの異常検知に与える影響を分析
- 時系列データのプライバシ保護手法が各異常の検知に及ぼす影響を調査

関連技術

Anomaly Transformer
Transformerを応用した時系列データの異常検知手法

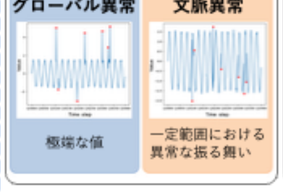


- Prior-Association
  「隣接データ点ほど関連性が強い」という事前知識のもと、隣接点の関連性を学習
- Series-Association
  Self-Attention同様、系列内の点と点の関連性を学習
- 関連性差異：2つのAssociationで定義

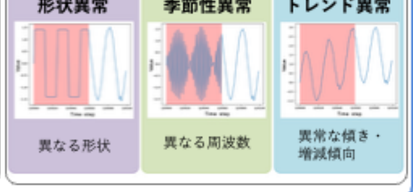
時系列データに含まれる異常
ポイント型2種類、集団型3種類の計5種類

ポイント型：個々のデータ点における異常な事象
集団型：複数のデータ点で構成、正常時と異なる挙動を示す部分時系列

ポイント型



集団型



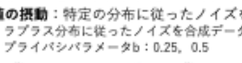
実験概要

使用データセット	種類	次元数	訓練データ数	テストデータ数	異常データの割合	含まれる異常
合成データセット	正弦波＋微量ノイズ	1	200,000	200,000	0.15, 0.20, 0.25, 0.30, 0.35, 0.40	各異常を1種ずつ含む(異常5種類×割合6通りの計30個)
SWaT	水処理	51	495,000	449,919	0.12	集団型(形状/トレンド)多数、集団型異常のサイズ大
PSM	産業システム	25	132,481	87,841	0.28	ポイント型(文脈)多数、集団型異常のサイズ小

実験結果

実験１　合成データセットに対する異常検知

異常の割合が高い場合において、**全ての異常を高い性能で検知可能**



季節性異常：異常の割合によらず**精度**が**高**、2つのAssociationにより時刻間の関連性を学習

グローバル異常・形状異常：異常の割合によらず**精度**が**高**、極端な値や形状の相違がAttentionにより容易に識別

トレンド異常：**異常の割合**が**大きく**なるほど**精度**が**向上**
データの値が正常時から大きく逸脱
→異常領域に挟まれた正常区間の多くが異常と判定
→異常領域間に位置する正常データの減少に伴い、偽陽性も減少

文脈異常：他の異常より精度が劣る
ポイント型 & 異常が**正常データの値域内に点在** → 識別難易度高

実験２　値の摂動を施した合成データセットに対する異常検知

値の摂動：特定の分布に従ったノイズをデータに付加
- ラプラス分布に従ったノイズを合成データに付加
- プライバシパラメータb：0.25，0.5

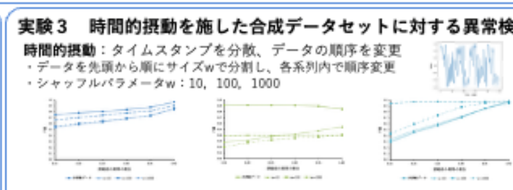
グローバル・形状異常
- ノイズの影響大
- ノイズ量の変化に敏感

文脈・季節性異常
- ノイズの影響大
- ノイズ量の変化に鈍感

トレンド異常
- ノイズの影響最小
- ノイズ量の変化に鈍感

実験３　時間的摂動を施した合成データセットに対する異常検知

時間的摂動：タイムスタンプを分散、データの順序を変更
- データを先頭から順にサイズwで分割し、各系列内で順序変更
- シャッフルパラメータw：10，100，1000



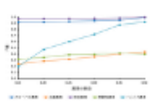
グローバル異常
摂動範囲によらず精度への影響小

文脈・形状・季節性異常
摂動範囲によらず影響大

トレンド異常
広範囲にデータが分散するほど精度向上

実験４　GAN vs Anomaly Transformer

合成データセット



- **GAN**（左図より）
  - データの値の分布に基づいて検知を実施
- **Anomaly Transformer**（実験１より）
  - Anomaly-Attentionを通じて周期構造や時刻間の関連性の変化を捉えることが可能に
  - データの値そのものへの依存はGANに比べ小さく、より構造的な検知を実施

実データセット

10回の検知を実施、平均値を算出

	MAD-GAN+Encoder	Anomaly Transformer
SWaT	0.643	0.647
PSM	0.464	0.552

Anomaly Transformerは正常データの値域内に分布する異常の検知に有効であり、複雑な異常を含む実データセットにおいてもその性能を実証

まとめと今後の課題

まとめ
- 検知対象のデータや含まれる異常の特徴に応じた検知手法の選択が重要
- データや検知手法の特性を踏まえたプライバシ保護手法の適応が不可欠

今後の課題
- 多様なデータ・手法での検証、保護手法の評価方法について検討
  →プライバシ保護を施したデータの高精度な検知手法を提案

インタラクティブなシーケンシャルパターンマイニングの性能向上に向けた提案と評価
（研究担当:青柳　結衣）

1.研究背景

- ビッグデータの蓄積に伴い、シーケンシャルパターンマイニング(SPM)が注目され、インタラクティブなSPMが不可欠
- 既存研究：本来生成されるべき候補シーケンスが生成されない場合があり、単調増加や増減の場合は実行時間がかかる

➡ 漏れなくシーケンスを生成しつつ、KB構造を見直すことでアルゴリズムの高速化を目指す

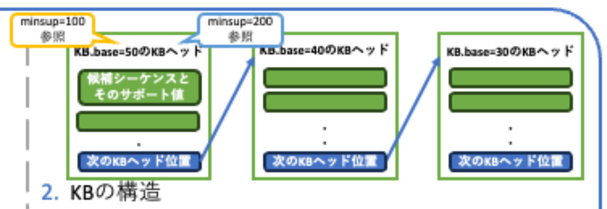
2.関連研究

KISP[1]

1. 直接計算による新しい候補シーケンス生成
   - 以下の式で候補シーケンス生成

$$X'_k = (S_{k-1}[KB.base] \otimes N_{k-1}[minsup]) \cup (N_{k-1}[minsup] \otimes N_{k-1}[minsup])$$

- X'_k：長さkの新しい候補シーケンス
- $S_{k-1}[KB.base]$：KB内に存在する長さk-1の頻出パターン
- $N_{k-1}[minsup]$：今回指定したminsupで新しく頻出になった長さk-1のパターン



2. KBの構造
   - 最小KB.baseより小さいminsupが指定された場合、新しいKBヘッド(候補シーケンスとサポート値)を作成

[1]Ming-Yen Lin and Suh-Yin Lee, "Improving the Efficiency of Interactive Sequential Pattern Mining by Incremental Pattern Discovery", Proceedings of the 36th Hawaii International Conference on System Sciences - 2003

3.提案手法

$$X'_k = (S_{k-1}[KB.base] \otimes N_{k-1}[minsup]) \cup (N_{k-1}[minsup] \otimes N_{k-1}[minsup]) \cup (N_{k-1}[minsup] \otimes S_{k-1}[KB.base])$$

生成式を追加

1. 生成式を変更し、全ての候補シーケンスを生成
   - 例：[f, g, c](サポート値2)の動き(minsup=5,2指定)
     - [f, g]：サポート値4、[g, c]：サポート値6
     - 既存手法では [f, g, c]が生成されない

2. KBの構造
   - 最大のKB.baseよりも大きいminsupを指定した場合も、新しくKBヘッドを作成

4.実験結果

minsup	6	8	10	9
シーケンス数	18,040	4,271	1,986	2,845

minsup	6(ms)	8(ms)	10(ms)	9(ms)
KISP	729,009.14	37.27	30.77	30.25
提案手法	721,008.69	3,029.30	1,309.31	29.98

1. minsup毎の実行時間の比較
   - minsupが増減する場合の実行時間を比較
     - BMSWebView1に適用し、平均実行時間を算出
   - ✓ minsupを増加させた後に減少させた最初の値であるminsup=9の実行時間は減少
     - ✓ シーケンス数の少ないKBヘッドを参照しているため、実行時間が短縮
     - ✓ 一方、KB作成時のminsup=8,10の実行時間は増加

2. 合計実行時間の比較
   - minsupが単調増加する場合と増減する場合の実行時間を比較
     - BMSWebView1とBMSWebView2に適用し、平均実行時間を算出
   - ✓ 単調増加時の実行時間差は僅か
     - ✓ 提案手法はKBヘッドの作成にコストがかかるが、KISPとほぼ同等
   - ✓ 増減する場合の実行時間は減少
     - ✓ 提案手法はminsupにより近い値のKB.baseを持つのKBヘッドから抽出できるため高速

単調増加時	KISP(s)	提案手法(s)	提案手法/KISP
BMSWebView1	127.66	126.51	99.10%
BMSWebView2	1,531.20	1,535.00	100.25%

増減時	KISP(s)	提案手法(s)	提案手法/KISP
BMSWebView1	129.02	126.03	97.68%
BMSWebView2	1,525.86	1,524.91	99.94%

5.まとめと今後の課題

まとめ
- ✓ 候補シーケンス生成式を変更し、すべての頻出シーケンスを抽出
- ✓ KB構造を見直すことで、minsupが増減する場合は実行時間が減少

今後の課題
- ✓ より大きいサイズのデータセットに適用し、評価
- ✓ KBヘッド作成のコストを抑えるため、KBヘッドの作成タイミングを調整

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