合成アルゴリズムの秘匿属性推定攻撃に対する安全性評価

谷口 輝海

研究背景

- データを活用したAIサービス等の普及 → データに機微な情報を含む
- 実データと**同じ統計的性質**を持つ, アルゴリズムで生成した**架空のデータ**
- 合成データのプライバシー ex) 属性推定

D_{orig}

学年	研究室	菊池論
3	斉藤研	A
3	斉藤研	A
4	斉藤研	F
3	菊池研	A

D_{syn}

学年	研究室	菊池論
3	斎藤研	A
4	菊池研	F
3	斎藤研	A
4	菊池研	A

推定に用いる変数　　推定したい変数

合成データで学習すればモデルは安全か?

- MLモデルの振る舞いから学習データを推論する攻撃が提案されている

D_{orig}

学年	研究室	菊池論
2	菊池研	F
3	斉藤研	F
4	菊池研	A
3	斉藤研	A

生成

D_{syn}

学年	研究室	菊池論
3	菊池研	A
2	斉藤研	F
3	斉藤研	A
4	菊池研	F

学習

MLモデル

推論

菊池論
F
F
A
A

研究目的

- 3つの合成アルゴリズムについてプライバシーリスクと有用性を定量化して比較する.
- **合成データで学習した機械学習モデルに, 推論攻撃への耐性があるか調査する.**

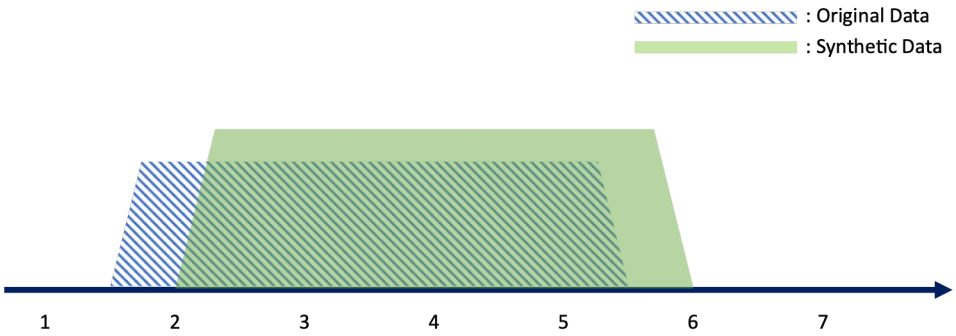
合成アルゴリズム

- **Conditional Tabular GAN (CTGAN)**
  - GANを表形式データ生成に応用したもの
- **Tabular Variational Auto Encoder (TVAE)**
  - VAEを表形式データ生成に応用したもの
- **CopulaGAN**
  - CTGANとCopulaの組合せ

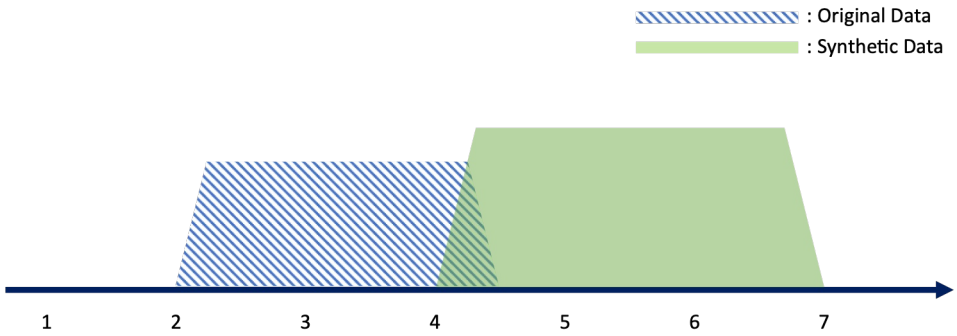
有用性評価 (Karr, 2012)

- Confidence Interval Overlap (CIO)
- 回帰モデルにおける信頼区間の重複度合いによる類似度評価
- 1に近いほど良い, 0に近いほど悪い

良い例



悪い例



リスク評価 (Taub, 2018)

- Targeted Correct Attribution Probability (TCAP)
- 合成データを用いたオリジナルデータの属性推定
- 1に近いほど高リスク, 0に近いほど低リスク

D_{orig}

学年	研究室	菊池論
2	菊池研	F
3	斉藤研	A
2	菊池研	A
3	斉藤研	A
4	斉藤研	F
4	菊池研	A

D_{syn}

学年	研究室	菊池論
3	斎藤研	A
2	菊池研	F
3	斎藤研	A
4	菊池研	F
2	菊池研	F
4	菊池研	A

key変数　　target変数

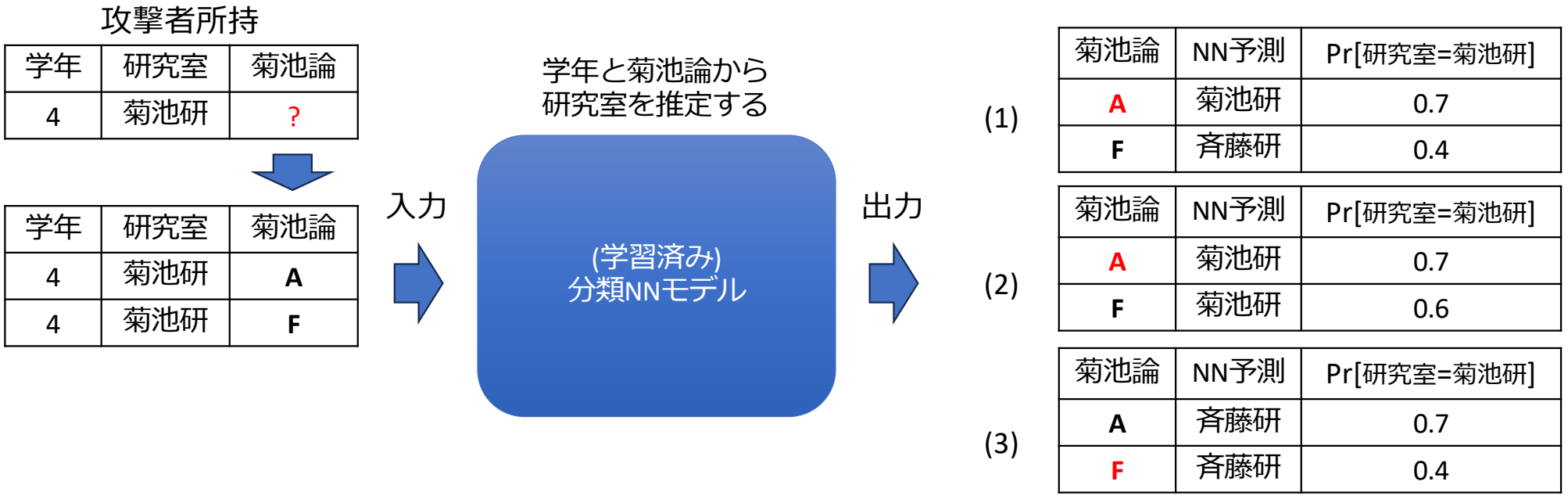
D_{syn}で, **(3, 斉藤研, A)**, **(2, 菊池研, F)** はkeyとtargetの組合せが一意.

D_{orig}でkeyが, **(3, 斉藤研)**, **(2, 菊池研)** の4レコード中, 菊池論の値が同じものは3つ.

TCAP = 3/4 = 0.75

MLモデルの属性推論攻撃 (Mehnaz, 2022)

- Confidence Score based Model Inversion Attack (CSMIA)
- 機械学習モデルの出力から学習データの属性(説明変数)を推定する.



実験

1. 信頼区間重複度合い(CIO)による有用性評価実験
2. 属性推定リスク(TCAP値)によるリスク評価実験
3. 合成データで学習したMLモデルに対する説明変数推定攻撃(CSMIA)の精度評価実験

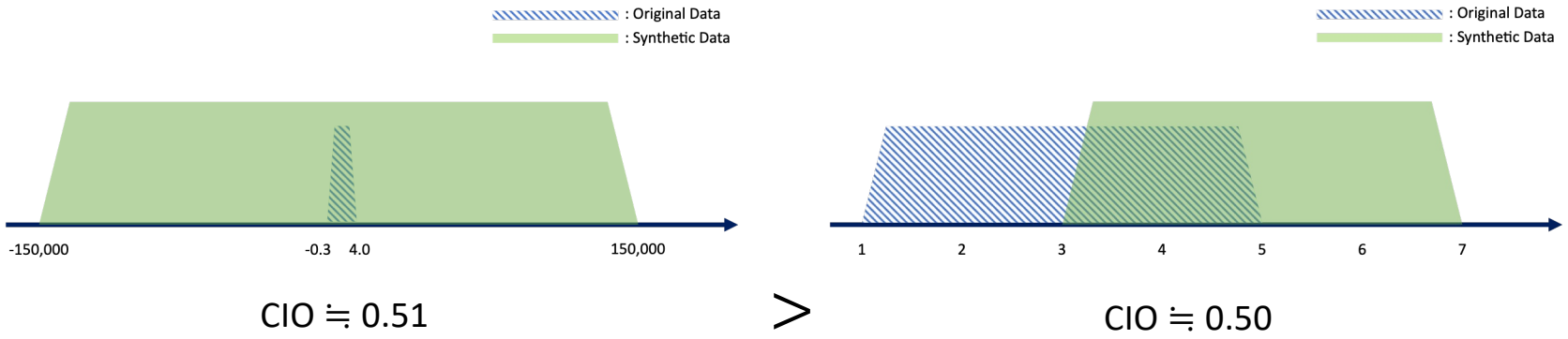
- 実験には, Adult Dataset を使用

実験1 結果 CIO

- age, workclass, education, hours.per.week, raceでincomeをlogistic回帰.

合成器	F1	最小値	中央値	最大値	平均値	標準偏差
オリジナル	0.66	-	-	-	-	-
CTGAN	0.64	0.00	0.62	0.92	0.56	0.31
TVAE	0.65	0.00	0.00	0.91	**0.16**	0.27
CopulaGAN	0.69	0.00	0.41	0.88	0.40	0.36

- 問題(TVAE)



CIO ≒ 0.51 > CIO ≒ 0.50

実験2 結果 TCAP

- workclass, relationship, race, marital.status, sex, income の中からkey変数とtarget変数を選び、TCAPを計算した結果の基本統計量

- TVAEで平均が高い値

- key変数
  - 少ない → CTGAN低リスク
  - 多い → CopulaGAN低リスク

合成器	key変数の数	組合せ数	最小値	最大値	平均値	標準偏差
CTGAN	3	60	0.000	1.000	**0.602**	0.363
	4	30	0.301	0.977	**0.762**	0.152
	5	6	0.694	0.939	0.798	0.108
TVAE	3	60	0.340	0.995	**0.810**	0.149
	4	30	**0.682**	0.970	**0.855**	0.089
	5	6	**0.800**	0.958	**0.890**	0.071
CopulaGAN	3	60	0.000	1.000	0.666	0.350
	4	30	0.462	0.970	0.779	0.136
	5	6	0.622	0.926	**0.786**	0.118

実験3 結果 CSMIA

- incomeを予測するNN
- marital.statusをMarriedとSingleの2値として推定
- CTGANで精度が低下

合成器	sensitive attribute	Precision	Recall	F1-Score	Accuracy (CSMIA)	Accuracy (NN)
オリジナル	Married	0.684	0.420	0.520	0.628	0.846
	Single	0.605	0.821	0.628		
CTGAN	Married	0.374	0.553	0.446	**0.340**	0.826
	Single	0.259	0.144	0.185		
TVAE	Married	0.645	0.458	0.536	0.619	0.813
	Single	0.605	0.768	0.677		
CopulaGAN	Married	0.694	0.410	0.515	0.630	0.825
	Single	0.605	0.833	0.701		

結論

- 3つの合成アルゴリズムCTGAN, TVAE, CopulaGANについて有用性とリスクを評価した.
- CIOによる有用性評価
  - TVAEが0.16で他の2つよりも大幅に低い結果となった
- TCAPによるリスク評価
  - 3つのうち、TVAEで高リスク
  - key変数が3~4つのときCTGAN, 5つのときCopulaGANで低リスク
- CSMIAによるリスク評価
  - CTGANで推定リスクが低下する