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 †

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室 †

1 はじめに

近年，データの利活用の活発化に伴い，データに含まれる個人のプライバシー保護が重要な課題となっている．データのプライバシーを保護する枠組みとして匿名化 [10] や差分プライバシー [9] といった技術が提案されている．しかし，これらの技術では，十分なプライバシーを提供するために，しばしばデータの有用性が著しく損なわれることがある．そこで，合成データ技術が注目されている．

合成データとは，実在するデータと同じ統計的性質を有するアルゴリズムで生成された架空のデータである．合成データはプライバシーを保護しながら実データに近い有用性を保つことができる手段として期待されているが，プライバシーの保護度合いは明確に定量化されておらず，どのような合成アルゴリズムにおいて有用性や安全性が高く保証されるかは定かではない．

そこで，本研究では Python のオープンソースライブラリとして Synthetic Data Vault[1] で提供されている，3 つの合成アルゴリズム (Conditional Tabular GAN[2], Tabular Variational Auto Encoder[2], CopulaGAN[3]) について，分布誤差，データ列間の相関の誤差，回帰モデルにおける信頼区間の重複度合いによる有用性評価指標で，生成される合成データの有用性を定量化する．次に，安全性について，(1) 合成データからの属性推論成功率を考慮したリスク評価指標と，(2)mehnaz らによって提案された，機械学習モデルの入出力の情報から学習データの属性値を推論する攻撃である Confidence Score based Model Inversion Attack[7] を，用いた開示リスクを評価する．CSMIA を，オリジナルのデータで学習したモデルと，合成データで学習したモデルのそれぞれに適用し，攻撃精度の差を明らかにすることを目的とする．

本研究のシステム構成図を図 1 に示す．このシステム構成図では，CSMIA を用いて合成データを評価するための手順を示している．まず，オリジナルデータセット D を合成アルゴリズムに入力し，合成データセット D' を得る．次に，D と D' を用いて分類ニューラルネットワーク f_{orig} と f_{synth} を学習させる．f_{orig} と f_{synth} を用いてオリジナルデータの変数 x_2 を推定する．続いて，f_{orig} と f_{synth} に，オリジナルデータのレコードを x_2 の可能な値で繰り返したデータを入力する．最後に，それぞれの出力に対して CSMIA を適用することで，x'_2 を推定する．

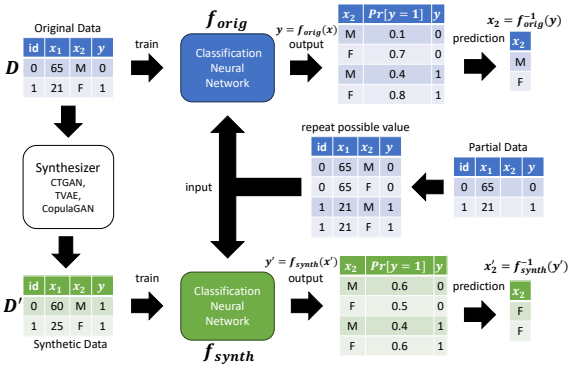

図 1　システム構成図

2 準備

2.1 Synthetic Data Vault

本研究では，datacebo 社が開発したオープンソースである Synthetic Data Vault ライブラリ [1] で提供される 3 つの合成アルゴリズム，CTGAN，TVAE，CopulaGAN を対象とする．

2.1.1 CTGAN

CTGAN[2] は，Generative Adversarial Network[11] の一種であり，表形式のデータを生成するように設計されている．通常の GAN と同様に，生成器 (Generator) と識別器 (Discriminator) と呼ばれる 2 つのニューラルネットワークを競合させ，同時に学習させることで，学習データに類似したデータを生成するモデルを構築する．

表形式データ生成における大きな課題点として，連続値属性の分布の複雑性と離散値であるカテゴリ出現頻度の不均衡の 2 つが挙げられる．CTGAN では，Mode-Specific Normalization と Training-by-Sampling という 2 つの手法を用いて表形式データ生成特有の問題点に対処

†Kikuchi Laboratory, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University.

している. Mode-Specific Normalization では，任意の複雑な分布を持つ連続値属性を，複数のガウス分布の線形和として表現し，分布のモードごとに数値を正規化する. Training-by-Sampling では，各カテゴリの対数頻度に応じて，識別器に入力するデータと，生成器に入力する条件ベクトルをサンプリングすることで，全ての取り得る離散値を均等に学習させることができる.

2.1.2 TVAE

Xu らによって提案された TVAE[2] は，表形式のデータを生成するために VAE を派生させたモデルである. CTGAN と同様の前処理を行い，通常の VAE と同じく学習を行う.

2.1.3 CopulaGAN

CopulaGAN[1, 3] は CTGAN と Copula[3] を組合わせた表形式データ生成モデルである. Copula は，多変量の累積分布関数と周辺分布関数の関係を表す関数であり，確率変数間の多様な依存関係を表現する.

2.2 相関値

表形式データでは，量的データと質的データが混在しているため，本研究では，データ形式に応じて，相関係数 (量的)，相関比 (量的と質的の組)，クラメルの連関係数 (質的) をそれぞれ相関の指標として用いる.

2.2.1 相関比

相関比 [6] は，量的変数と質的変数の関係の強さを表す指標である. 0 から 1 の値を取り，1 に近いほど変数間の関連が強い.

C を取り得るカテゴリの集合とし，各カテゴリの変数の個数を n_i 個 ($i \in C$) とする. また，$x_{i,j}$ をカテゴリ i であるもののうち，j 番目の量的変数値とし，量的変数の平均値を $\overline{x}$，質的変数の各カテゴリ毎の i 番目の量的変数の平均値を $\overline{x_{i,j}}$ とする. このとき，相関比 η^2 は以下の式で計算される.

$$\eta^2 = \frac{\sum_{i \in C} n_i (\overline{x}_i - \overline{x})^2}{\sum_{i \in C} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2}$$

表 2 の例における **age**(量的変数) と **marriage**(質的変数) を用いて，

$\overline{x_{age}} = 40.8, \overline{x}_{age,Single} = 34,$

$\overline{x}_{age,Marital} = 47, \overline{x}_{age,Divorced} = 42$

のとき，

$$\eta^2 = \frac{2 \cdot (\overline{x}_{Single} - \overline{x})^2 + 2 \cdot (\overline{x}_{Marital} - \overline{x})^2 + 1 \cdot (\overline{x}_{Divorced} - \overline{x})^2}{(43 - \overline{x})^2 + (29 - \overline{x})^2 + (25 - \overline{x})^2 + (42 - \overline{x})^2 + (65 - \overline{x})^2} \fallingdotseq 0.066$$

である.

2.2.2 クラメルの連関係数 [8]

クラメルの連関係数は質的変数同士の関連の強さを表す指標である. 0 から 1 の値を取り，1 に近いほど関連が強いとされる. 質的変数のカテゴリの数をそれぞれ r, c としたとき，r 行 c 列のクロス集計表を作成し，その χ^2 値から算出される. サンプルレコード数を n とすると，連関係数 V は，

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

で定義される. 表 2 の例における **marriage**(質的変数) と **sex**(質的変数) を用いた計算例を示す. サンプルデータにおけるクロス集計表は表 1 のような 3×2 の表となる. ここから χ^2 値を算出すると，$\chi^2 \fallingdotseq 3.3$ となり，

$$V = \sqrt{\frac{3.3}{5 \cdot \min(3-1, 2-1)}} = 0.66$$

表 1　D のクロス集計

sex / **marriage**	**Male**	**Female**	total
Single	2	1	2
Marital	0	2	2
Divorced	1	0	1
total	3	2	5

表 2　サンプルデータ D

id	**age**	**height**	**marriage**	**sex**	**income**
0	43	170	Single	Female	$> 50K$
1	29	161	Marital	Female	$\leq 50K$
2	25	179	Single	Male	$\leq 50K$
3	42	174	Divorced	Male	$> 50K$
4	65	153	Marital	Female	$> 50K$

2.3 Confidence Interval Overlap [4]

合成データの有用性を測るために，Karr らによって提案されている Confidence Interval Overlap (CIO) を使用する. オリジナルデータ D と合成データ D' でそれぞれ構築した回帰モデルにおいて，回帰係数の信頼区間に占める重複区間の割合の平

表 3　合成サンプルデータ D'

id	age	height	marriage	sex	income
0	40	175	Single	Male	$\leq 50K$
1	35	155	Single	Male	$> 50K$
2	25	171	Marital	Female	$\leq 50K$
3	42	165	Divorced	Female	$> 50K$
4	52	145	Marital	Female	$\leq 50K$

均で計算される. u_o, ℓ_o と u_s, ℓ_s をそれぞれ，オリジナルデータ D と合成データ D' に対する回帰係数の信頼区間の上界と下界とするとき，CIO は，

$$CIO = \frac{1}{2}\left(\frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_o - l_o} + \frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_s - l_s}\right)$$

と定める. 信頼区間が完全に一致する場合，CIO は最大値 1 をとる. 本研究では重複がない場合の CIO 値を 0 とする.

例えば，あるデータについて，図 2 に示すようにオリジナルデータが [1, 5]，合成データが [3, 7] となる信頼区間が与えられたとする. このとき，CIO の値は，

$$\begin{aligned} CIO &= \frac{1}{2}\left(\frac{\min(5,7) - \max(1,3)}{5-1} + \frac{\min(5,7) - \max(1,3)}{7-3}\right) \\ &= 0.5 \end{aligned}$$

である.

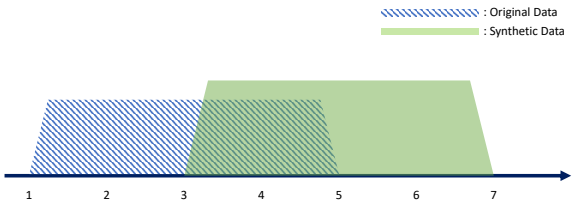

図 2　信頼区間の例

2.4　Targeted Correct Attribution Probability [5]

Taub らによって提案された Targeted Correct Attribution Probability(TCAP) は，合成データを公開したときに元データの値がどれだけ漏洩するかのリスクを評価するための指標である. TCAP の算出において，敵対者は合成データにアクセスできると仮定し，質的変数をいくつか得る. 1 つを target 変数，残りを target 変数を予測するための key 変数とする. オリジナルデータと合成データにおける key 変数と target 変数をそれぞれ K_o, T_o, K_s, T_s とする. このとき，TCAP を計算するために，合成データの各レコード j について，レコード j がもつ K_s の組合わせと同じレコードのうち，target 変数がレコード j と一致するものの割合を算出する. これは，Within Equivalece Class Attribution Probability (WEAP) と呼ばれ，

$$\begin{aligned} WEAP_{s,j} &= \Pr(T_{s,j} | K_{s,j}) \\ &= \frac{Pr[T = T_{s,j}, K = K_{s,j}]}{Pr[K = K_{s,j}]} \\ &= |\{(K_{s,i}, T_{s,i}) \mid T_{s,i} = T_{s,j}, K_{s,i} = K_{s,j}, i \in [n]\}| \end{aligned}$$

と定義される.

$WEAP$ の値を閾値として, TCAP を計算するレコードを定める. 本研究では WEAP の値を 1 とした. つまり，$WEAP_{s,j} = 1$ であるような合成レコード j について TCAP の値を算出する. 対応するオリジナルデータのレコード j について，TCAP を.

$$\begin{aligned} TCAP_{o,j} &= \Pr(T_{s,j} | K_{s,j})_o \\ &= \frac{\sum_{i=1}^{n}[T_{o,i} = T_{s,j}, K_{o,i} = K_{s,j}]}{\sum_{i=1}^{n}[K_{o,i} = K_{s,j}]} \\ &= \frac{Pr[T = T_{o,j}, K = K_{s,j}]}{Pr[K = K_{s,j}]} \end{aligned}$$

とする.

$WEAP_{s,j} = 1$ に対応するオリジナルレコードが存在しない場合，分母は 0 となるため TCAP は定義されない. TCAP は 1 に近いほど開示リスクが高く，0 に近いほど開示リスクは低い.

表 2 のデータセットを用いて表 3 のような合成データセットが得られたとする. $K_s = \{\textbf{marriage}, \textbf{sex}\}, T_s = \{\textbf{income}\}$ としたとき，合成データにおいて key 変数の組合せは $id = 0, 1$ が (*Single*, *Male*), $id = 3$ が (*Divorced*, *Female*), $id = 2, 4$ が (*Marital*, *Female*) となり，3 パターン存在する. ここで，$id = 0, 1$ については target 変数の値がそれぞれ異なるため $WEAP_{s,()} = 0.5$ である. $id = 2, 3, 4$ については target 変数の値が一意に定まっているため，$WEAP = 1$ となる. $WEAP = 1$ となるレコードの key 変数に一致するレコードをオリジナルデータ D(表 2) から抜き出すと，$id = 1, 4$ が該当する. このうち，target 変数が一致するレコードは $id = 1$ であるので，TCAP の値は 0.5 となる.

2.5　Confidence Score based Model Inversion Attack [7]

Mehnaz らによって提案された Confidence Score based Model Inversion Attack(CSMIA) は，学習済み分類モデルの出力値から学習データの属性推論を行うアルゴリズムである. 攻撃者は，標的モデルに入力を行うことができ，分類結果と各クラスの所属確率にアクセスできるものとする. 分類モデルの学習に用いたデータを $x_1, x_2, \ldots, x_d, y$ とする. ただし，x_i は説明変数，y は目的変数 (クラスラベル) である. このとき，攻撃者は $x_2, \ldots, x_d, y$ に関する情報を有しており，あるセンシティブな属性 x_1 について推論を行う.

CSMIA では，モデルに対し正しい属性値で入力を行なった場合に, 矛盾のない, より高い確信度で正しい予測結果を返すという考えに基づき, 属性を推論する. まず, 攻撃者はセンシティブ属性について可能な値を全て列挙した入力用のデータを作成し，モデルに対して入力を実行する. 例えば，x_1 が **Marital** であるとき，考えられる値は Married と Single の 2 つであり，攻撃者は機械学習モデルに対して，$(Married, x_2, \ldots, x_d)$ と $(Single, x_2, \ldots, x_d)$ を入力し，出力 $\hat{y}_{Married}$ と $\hat{y}_{Single}$ を得る. そして，以下の 3 ケースについて，センシティブ属性値を推論する.

(1) $\hat{y}_{Married} = y$ かつ $\hat{y}_{Single} \neq y$，もしくは，$\hat{y}_{Married} \neq y$ かつ $\hat{y}_{Single} = y$ のとき，y について正しい出力をした入力を推論値とする. 例えば，$\hat{y}_{Married} = y$ かつ $\hat{y}_{Single} \neq y$ であれば $\hat{x_1} = Married$ を返す.

(2) $\hat{y}_{Married} = y$ かつ $\hat{y}_{Single} = y$ のとき
y の所属確率が高い方を推論値とする.

(3) $\hat{y}_{Married} \neq y$ かつ $\hat{y}_{Single} \neq y$ のとき
y の所属確率が低い方を推論値とする.

3 実験方法

3.1 実験目的

SDV ライブラリで提供される 3 つの合成アルゴリズムの有用性と開示リスクを比較評価するために，様々な評価指標を用いて実験を行う.

3.2 使用データ

本研究では, Adult データセット [12] を用いた. 表 4 に Adult データセットの概要を示す.

表 4　Adult データセットの概要

変数名	データタイプ	カテゴリ数	詳細
age	連続	-	年齢
workclass	質的	7	職業クラス
fnlwgt	連続	-	重み（Final Weight の略）
education	質的	16	教育レベル
marital.status	質的	7	結婚の状態
occupation	質的	14	職業
relationship	質的	6	家族関係
race	質的	5	人種
sex	質的	2	性別
capital.gain	連続	-	資産利益
capital.loss	連続	-	資産損失
hours.per.week	連続	-	週の労働時間
native.country	質的	41	出身国
income	質的	2	収入（$\leq 50K, > 50K$）

3.3 有用性評価

3.3.1 分布評価

データセットの属性ごとに頻度分布を求め，オリジナルデータと合成データとの間で各カテゴリ値の出現確率の誤差を求めた. ただし，数値変数に関しては，階級を 16 個に分割して計算した. 誤差の算出には Mean Abusolute Error(MAE) を用いる.

3.3.2 相関評価

データセット内の各属性の組について相関値を計算し，オリジナルデータと合成データとの間で MAE を求めた. なお，データセット内には量的変数と質的変数が混在しているため，組合せる属性のデータ形式に応じて相関指標を選択している. 量的変数間に相関係数, 量的変数と質的変数に相関比, 質的変数間にクラメルの連関係数を用いた.

3.3.3 CIO

オリジナリデータと合成データの類似度を CIO を用いて評価する. ロジスティック回帰を使用し，各回帰係数に対する CIO の基本統計量を算出する. 説明変数には age，workclass, education，hours.per.week，race を使用し，目的変数は income とする.

3.4 リスク評価

3.4.1 TCAP

各合成アルゴリズムで生成した合成データについて TCAP の値を算出し，開示リスクを評価する.

表 5 に各合成データで TCAP 値算出のために用いた変数を示す. key 変数の数を 3 ~ 5 個とし，key 変数と target 変数の全組合せ 96 通りについて TCAP 値を算出した.

表 5　使用変数とカテゴリの数

変数	workclass	relationship	race	marital.status	sex	income
カテゴリ数	7	6	5	7	2	2

3.5 CSMIA

オリジナルデータで学習した機械学習モデルと，オリジナルデータを用いて生成した合成データで学習した機械学習モデルに対し，CSMIA を実行し，機械学習モデルの精度と攻撃の精度の関係を調査する.

攻撃対象のモデルには, income の 2 値分類を行うニューラルネットワーク (NN) をオープンソースライブラリ PyTorch[13] で実装した. NN は 3 層の全結合層で構成され, 中間層の活性化関数には ReLU 関数，出力層の活性化関数には sigmoid 関数を用いた. 最適化アルゴリズムは Adam を採用し, 学習率は 0.005 としている. また，推論対象の変数を marital.status とし，7 つのカテゴリ値の内，Married-civ-spouse, Married-spouse-absent, Married-AF-spouse を Married，Divorced, Never-married, Separated, Widowed を Single としてまとめて 2 値変数とした.

4 実験結果

4.1 有用性評価

4.1.1 頻度分布

図 3 に各属性のカテゴリ値出現確率の平均絶対誤差 (MAE) を示す. ここで，CTGAN の値でソートしている.

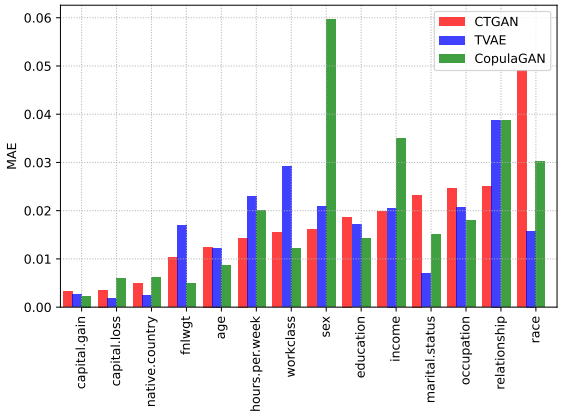

図 3　分布誤差

4.1.2 相関評価

図 4 に各合成器の相関指標ごとの相対絶対誤差 (MAE) を示す. ここで，CTGAN の値でソートしている.

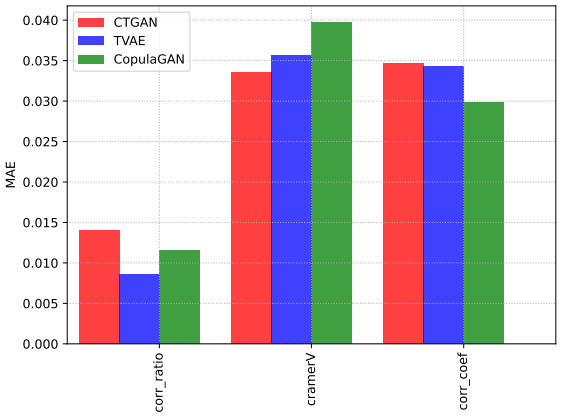

図 4　平均相関誤差

4.1.3 CIO

表 7 に各合成データで算出した回帰係数の上限と下限の値を示す. 表 7 における Overlap は，1 つの変数から算出した回帰係数における CIO の値である. 表 6 に各合成データとオリジナルデータとの間で計算した CIO の結果を示す. なお，CIO を計算するために使用したロジスティック回帰の F1-score を表中に示している.

表 6　CIO 値の基本統計量

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

4.2 リスク評価

4.2.1 TCAP

key 変数を 3 ~ 5 個としたときの全組合せについて，TCAP 値を算出した. 表 8 に各合成器ごとの TCAP 値の基本統計量を示す.

4.3 CSMIA

実験に使用した NN の分類精度を表 9 に示す. なお，合成データで学習した機械学習モデルのテストデータには，オリジナルデータを分割して作成したテストデータを使用している. 各データで学習した NN に対して，CSMIA を実行した結果を表 10 に示す. ここで，CSMIA においては，オリジナルモデルの学習データについて推論をした. 評価指標として Precision, Recall, f1-Score, Accuracy を用いた.

4.4 考察

4.4.1 分布/相関評価

図 3 で，各合成器において，最も誤差が小さくなった変数の数は CTGAN が 4 つ，TVAE が 4 つ，CopulaGAN が 6 つとなった. income や sex といった 2 値変数では CopulaGAN の誤差が顕著に大きくなっている. また，CTGAN において最も誤差が小さい変数では，カテゴリの数が少ない変数が多く，CopulaGAN において最も誤差が小さい変数では，カテゴリの数が多い変数が多い. 一方，最もカテゴリ数が多い native.country においては TVAE が誤差が最も小さく，カテゴリ出現頻度誤差についての，一貫した誤差の傾向はないように考えられる.

図 4 の相関比と相関係数について，CTGAN における誤差が，他の合成器に比べて大きい. 一方で，Cramer の連関係数については，CTGAN のにおける誤差が最も小さくなっており，カテゴリ変数間の相関を保持している. また，相関係数では CopulaGAN，相関比では TVAE の誤差が最も小さくなっている. このことから，相関保持の観点からは，CTGAN や CopulaGAN といった GAN をベースとした合成方式ではカテゴリ変数の合成に適しており，TVAE では数値変数の合成に適している.

ただし，元々の相関の値が小さく，誤差の値も小さいため，大きな差は見られない.

4.4.2 CIO

CTGAN において，中央値，最大値，平均値について，他の 2 つの合成器と比べて高い値を示している. 最小値に関してはいずれの合成器においても 0.00 を示しているが，表 7 の各合成器における Overlap の値を見ると，CIO の値が 0.00 となる変数の数が，CTGAN では 4 個，TVAE では 16 個，CopulaGAN では 9 個となり，CTGAN 以外の合成器では多くの変数において，オリジナルデータと信頼区間が全く重複しておらず，類似性が低い.

表 7 で，TVAE の *workclass(T.Without-pay)*，*education(T.Assoc-voc)*，*education(T.Preschool)* において，信頼区間の幅が非常に広くなっている. このとき，オリジナルデータの信頼区間を全て包含し，CIO の定義式の第一項の値が 1 となるため，結果としては CIO が 0.5 を超えるものの，オリジナルデータと全く類似していないという問題が発生する. この問題に対処するためには，信頼区間の幅の比を用いて重み付けすることなどが考えられる.

また，TVAE において，表 7 の *education(T.Assoc-acdm)*，*race(T.Asian-Pac-Islander)* といった変数で，カテゴリが存在せず，これらの変数について CIO が算出できない場合がある. このようにサンプリング時にカテゴリの多様性が失われたことも，CIO を下げる一因であると考えられる.

CIO では，最大値，中央値，平均値において最も高い値を示した CTGAN がオリジナルデータを最もよく近似できている. 一方で，TVAE は，表 6 の太字に示すように，平均値において著しく低い値を示しており，標準偏差も小さいことから，オリジナルデータをうまく近似できていない. また，CIO では，信頼区間が完全に包含される場合に必ず 0.5 以上の値を示してし

表 7　各合成器における回帰係数毎の CIO

	orig_lower	orig_upper	CTGAN_lower	CTGAN_upper	TVAE_lower	TVAE_upper	CopulaGAN_lower	CopulaGAN_upper	CTGAN_Overlap	TVAE_Overlap	CopulaGAN_Overlap
Intercept	5.856	6.809	4.696	5.607	2.495	3.730	4.166	4.929	0.000	0.000	0.000
workclass(T.Local-gov)	0.447	0.777	0.339	0.751	2.387	3.501	0.998	1.337	0.829	0.000	0.000
workclass(T.Private)	0.380	0.655	0.293	0.594	2.077	3.172	0.848	1.116	0.747	0.000	0.000
workclass(T.Self-emp-inc)	−0.305	0.061	−0.348	0.087	0.382	1.593	0.221	0.589	0.921	0.000	0.000
workclass(T.Self-emp-not-inc)	0.713	1.039	0.454	0.796	2.614	3.725	0.940	1.244	0.249	0.000	0.316
workclass(T.State-gov)	0.607	0.978	0.546	0.935	2.532	3.751	1.395	1.743	0.864	0.000	0.000
workclass(T.Without-pay)	−0.320	3.938	2.171	3.996	**−155720.310**	**155765.088**	0.527	2.255	0.692	0.500	0.703
education(T.11th)	−0.556	0.175	0.320	1.390	1.485	2.570	0.146	0.815	0.000	0.000	0.042
education(T.12th)	−1.026	−0.130	−0.764	0.361	−2.448	0.746	−0.665	0.036	0.636	0.640	0.680
education(T.1st-4th)	−0.010	1.714	−0.471	0.606	0.158	1.585	0.098	1.156	0.465	0.914	0.807
education(T.5th-6th)	−0.173	0.932	−0.695	0.368	0.382	1.109	0.055	0.984	0.499	0.627	0.869
education(T.7th-8th)	−0.040	0.790	−0.340	0.442	1.005	1.647	−0.020	0.491	0.598	0.000	0.807
education(T.9th)	−0.123	0.843	0.047	1.011	0.802	2.293	−0.031	0.540	0.825	0.035	0.796
education(T.Assoc-acdm)	−2.107	−1.506	−1.653	−0.964	–	–	−1.755	−1.272	0.228	–	0.464
education(T.Assoc-voc)	−2.028	−1.441	−2.254	−1.573	**−62471.400**	**62515.005**	−1.783	−1.368	0.721	0.500	0.704
education(T.Bachelors)	−2.713	−2.165	−2.680	−2.048	−0.982	−0.628	−2.231	−1.840	0.877	0.000	0.145
education(T.Doctorate)	−3.777	−3.076	−3.451	−2.673	−2.096	−1.618	−3.562	−3.068	0.509	0.000	0.839
education(T.HS-grad)	−1.369	−0.823	−1.321	−0.688	0.259	0.615	−0.767	−0.374	0.849	0.000	0.000
education(T.Masters)	−3.069	−2.500	−3.229	−2.565	−1.998	−1.567	−2.240	−1.812	0.823	0.000	0.000
education(T.Preschool)	−0.531	3.497	−0.907	3.115	**−24382.939**	**24426.347**	−0.920	0.842	0.906	0.500	0.560
education(T.Prof-school)	−3.886	−3.226	−4.189	−3.464	−2.541	−1.959	−3.808	−3.303	0.611	0.000	0.883
education(T.Some-college)	−1.756	−1.206	−1.321	−0.665	0.082	0.442	−1.108	−0.716	0.192	0.000	0.000
race(T.Asian-Pac-Islander)	−0.791	−0.045	−0.717	−0.150	–	–	−1.158	−0.564	0.880	–	0.343
race(T.Black)	−0.277	0.434	−0.012	0.554	0.359	0.754	−0.139	0.469	0.707	0.149	0.874
race(T.Other)	−0.647	0.390	0.000	0.767	0.158	2.539	−0.202	0.520	0.443	0.160	0.696
race(T.White)	−0.896	−0.215	−1.103	−0.572	−0.296	−0.018	−1.235	−0.670	0.543	0.206	0.365
age	−0.049	−0.044	−0.020	−0.015	−0.041	−0.036	−0.027	−0.022	0.000	0.000	0.000
hours_per_week	−0.043	−0.038	−0.036	−0.031	−0.066	−0.059	−0.041	−0.036	0.000	0.000	0.581

表 8　各合成データにおける TCAP 値の基本統計量

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

表 9　NN による目的変数の識別精度

学習データ	y (income)	Precision	Recall	f1-Score	Accuracy
オリジナル	≤ 50K	0.875	0.928	0.900	0.846
	>50K	0.731	0.597	0.657	
CTGAN	≤ 50K	0.851	0.932	0.890	0.826
	>50K	0.711	0.506	0.591	
TVAE	≤ 50K	0.848	0.915	0.880	0.813
	>50K	0.662	0.502	0.571	
CopulaGAN	≤ 50K	0.875	0.895	0.885	0.825
	>50K	0.658	0.612	0.634	

表 10　CSMIA による属性推定精度

合成器	sensitive Attribute	Precision	Recall	f1-Score	Accuracy
Original	Married	0.684	0.420	0.520	0.628
	Single	0.605	0.821	0.628	
CTGAN	Married	0.374	0.553	0.446	**0.340**
	Single	0.259	0.144	0.185	
TVAE	Married	0.645	0.458	0.536	0.619
	Single	0.605	0.768	0.677	
CopulaGAN	Married	0.694	0.410	0.515	0.630
	Single	0.605	0.833	0.701	

まうことを考慮に入れて評価することが重要である.

4.4.3　TCAP

CTGAN と CopulaGAN では最小値，最大値，で顕著な差は見られないが，平均値では，key 変数の数が 3,4 つのときには，CopulaGAN の TCAP 値が CTGAN よりも高く，key 変数の数が 5 つになると，CTGAN の値が CopulaGAN の値を上回る. また，表 8 の TVAE の行の太字に示すように，TVAE では平均値と最小値において，他の 2 つの合成器に対して，高い値を示した. TVAE における TCAP の最小値を与えた変数の組は，key 変数が *race*，*marital.status*，*sex*，*income* 及び，*relationship*，*race*，*marital.status*，*sex*，*income* で，target 変数はいずれも *workclass* である. 平均的に見ると，TVAE が最もリスクが高く，key 変数の数が少ない時には CopulaGAN のリスクが高くなると考えられる.

4.4.4　CSMIA

NN の精度では，CTGAN が正解率において他の合成器と比べて高い精度を示し，CTGAN と CopulaGAN の一部指標においてオリジナリデータで学習した NN よりも高い結果となった.

しかし，CSMIA の精度においては，CTGAN で著しく推論精度が低下しており，CopulaGAN ではオリジナルよりも高い正解率で推論に成功している.

5　結論

本研究では，SDV ライブラリで提供される 3 つの表形式データ合成アルゴリズムである CTGAN，TVAE，CopulaGAN について，データ分布や相関の誤差，CIO や TCAP といった評価指標を用いて合成データの有用性とリスクを定量的に評価した. また，学習済み機械学習モデルの入出力から学習データの属性推論を行う CSMIA を用い，合成データで学習した機械学習モデルにおいて，どれだけ属性推論リスクがあるのかを

実験により明らかにした.

その結果，相関指標では，CTGAN がカテゴリカル変数間の相関保持に適しており，CopulaGAN では数値変数間の相関保持に適していることが分かった．また，CIO を用いた評価では，TVAE において著しく低い値を示し，オリジナルデータとの類似度が低いことが分かった. TCAP においては，key 変数が 3,4 つのとき，CTGAN で低く，key 変数が 5 つのとき CopulaGAN で低い値となった. さらに，TVAE では他の二つの合成器と比べて，TCAP の平均値が key 変数の数に関わらず 0.8 以上の値を示し，リスクが高いことが分かった. CSMIA を用いた評価では，CTGAN において属性推定リスクが大幅に軽減できることを実験的に示した.

しかし，CIO の計算において，一方のデータにおける信頼区間の幅が極端に大きくなってしまった場合に，CIO 値が高く出てしまうなど，有用性の指標としての課題があると考えられ，今後の研究課題としたい.

参考文献

[1] Patki, Neha and Wedge, Roy and Veeramachaneni, Kalyan, “The Synthetic data vault”, IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp.399-410, 2016.

[2] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni, “Modeling Tabular Data Using Conditional GAN”. Neural Information Processing Systems, pp.7335–7345, 2019.

[3] Yi Sun, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni, “Learning vine copula models for synthetic data generation”, AAAI Conference on Artificial Intelligence, pp.5049–50572018.

[4] Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P., ”A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality”. The American Statistician, pp.224-232, 2012.

[5] Jennifer Taub, Mark Elliot, Maria Pampaka, Duncan Smith, “Differential Correct Attribution Probability for Synthetic Data”, An Exploration, PSD 2018, p0.122-37, 2018.

[6] Ronald A. Fisher, “STATISTICAL METHODS FOR RESEARCH WORKERS”, 1925.

[7] Shagufta Mehnaz, Sayanton V. Dibbo, Ehsanul Kabir, Ninghui Li, Elisa Bertino, “Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models”, USENIX Security, pp.4579-4596, 2022.

[8] Cramér, H, “Mathematical Methods of Statistics”, Princeton University Press, pp.282. 1946.

[9] C. Dwork, “Differential privacy”, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, (eds) Automata, Languages and Programming, volume.4052, 2006.

[10] Sweeney, L., “k-anonymity: a model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, pp.557-570, 2002.

[11] I. J. Goodfellow, et al., “Generative Adversarial Nets”, Neural Inf. Process Syst, pp2672-2680, 2014

[12] Becker Barry, and Kohavi Ronny. 1996. Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

[13] Paszke, Adam and Gross.“Automatic differentiation in PyTorch”, NIPS-W, 2017.