F-041

# 深層生成モデルを用いた表形式データの拡張とその評価

## Augmenting Tabular Data by Deep Generative Models and Its Evaluation

小長谷 佳紀 [1] 英 彰吾 [2] 亀谷 由隆 [2]
Yoshiki Konagaya Shogo Hanabusa Yoshitaka Kameya
髙橋 和男 [3] 坪井 直毅 [3] 水野 智博 [3]
Kazuo Takahashi Naoki Tsuboi Tomohiro Mizuno

## 1 はじめに

機械学習で問題が生じる状況としてデータの不足が挙げられる．例えば医療分野ではデータの大規模な収集が困難な場合が多い．そこで画像認識分野を中心に，少ないデータ数を補うためにデータ拡張 (data augmentation) が一般的に行われている．データ拡張においては訓練データの画像にシフト・回転等の単純な加工を施したものを追加することが多いが，近年では深層生成モデルを用いたデータ拡張も行われており，医用画像にも試みられてる [5, 9]．その一方，データのプライバシー保護に関する研究分野 [13] では，表形式の元データと分布が近い人工データを生成する方法が近年提案されている [2, 8, 11]．

本研究では，表形式のデータに基づく分類モデルの学習において深層生成モデルを用いたデータ拡張を行うことを提案し，その有効性を評価する．対象とするデータは抗がん剤シスプラチンに起因する急性腎障害の発症予測 [7, 12] に関するものである．このデータに対し，我々はまず表形式データの生成モデルとして敵対的生成ネットワーク (generative adversarial network, GAN) をベースとする CTGAN [11]，変分オートエンコーダ (variational autoencoder, VAE) をベースとする TVAE [11] を利用することを考える．そして各々のモデルを使ってデータ拡張を行い，拡張後のデータから学習した分類分類モデルの分類精度が向上するかどうかを評価する．

評価においては，分類精度による定量的評価に加えて，機械学習モデルの判断根拠を示す説明可能 AI 手法の一つである SHAP [6] を用いることにより，生成されたデータの妥当性を評価する．更に，本研究では生成したデータ単独から学習した分類モデルの評価も行う．これは（CTGAN, TVAE の元々の目的である）プライバシー保護の観点から，データに個人情報が含まれていないときの分類精度を評価していることに相当する．

本論文では以下の構成をとる．はじめに 2 節では本研究で対象とするデータについて記述する．そして 3 節で提案手法，4 節で生成されたデータの評価方法を説明する．その後 5 節にて評価実験の結果を示す．6 節で論文のまとめを行い，今後の課題を述べる．

## 2 準備

本研究では 2006 年から 2013 年にかけて藤田医科大学病院で記録された非 ICU 患者データセットを対象とする [4]．このデータセットはシスプラチン (cisplatin) と呼ばれる抗がん剤の副作用として急性腎障害（シスプラチン誘発性急性腎障害, Cisplatin-induced acute kidney injury) が発生したかどうかを記録したものである [7, 12]．データセットはシスプラチン初回投与時の患者データのみが含まれている．

患者の特徴量として性別 (sex, 男性: 1, 女性: 0)，体表面積 (BSA [$m^2$])，年齢 (age [歳])，シスプラチンの 1 日最大投与量 (Cisplatin dose [mg/day])，血清クレアチニン値 (base sCr [mg/dl])，血清アルブミン値 (baseline albumine [g/dl])，糖尿病既往歴 (DM, あり: 1, なし: 0)，心血管イベント既往歴 (CVD, あり: 1, なし: 0) の 8 つが存在し，我々はこれらの特徴量から CTCAE グレード [5] が 1 以上になるかどうかを予測する分類タスクを考える．以降では CTCAE グレードが 1 以上という分類クラスを「陽性」，そうでない分類クラスを「陰性」と呼ぶ．これらの特徴量の中で，シスプラチンの 1 日最大投与量は人為的に制御できるため，医療的な観点から重要である [12]．

後述する本研究の評価実験では先行研究 [7, 12] と同様に，2006 年から 2012 年に記録されたデータ 1,014 件を訓練データ，2013 年に記録されたデータ 226 件をテストデータとして分割した．提案手法においては前者から人工の患者データを生成することを考えるため，前者を原 (raw) データと呼ぶことにする．訓練データにおける陽性患者は 184 件 (18.15%)，テストデータにおける陽性患者は 28 件 (12.39%) と分類クラス分布に偏りがある．本研究で考える人工データの生成は分類クラスの分布が偏るデータセットで発生しやすい少数派（今回は陽性）クラスの絶対的な希少性 (absolute rarity) [10] を緩和できる可能性がある．

## 3 提案手法

本研究では，表形式のデータに基づく分類モデルの学習において，元々はプライバシー保護の目的で提案されていた深層生成モデルを用いてデータ拡張を行うことを提案する．利用する深層生成モデルは先述の CTGAN, TVAE を利用する．すなわち，CTGAN もしくは TVAE を原データから学習し，生成した人工データを元の原データに加えて分類モデルの訓練データとする．

すなわち，分類モデルの学習に用いられるデータは原データ，CTGAN が生成するデータ，TVAE が生成するデータの 3 種類である．CTGAN, TVAE が生成するデータ数は原データの $k$ 倍 ($k = 1, 5, 10, 20$) とし，生成されたデータをそれぞれ CTGAN-$k$, TVAE-$k$ と参照する．原データは Raw と参照することがある．

そして本研究では，CTGAN, TVAE のどちらを使うか，どの程度の量の人工データを生成すればよいかを確認するための比較実験を行う．また，元々のプライバシー保護の観点から，原データは用いずに人工データのみから分類モデルを学習する場合も考える．

CTGAN [11] は，ランダムなノイズベクトルを入力として表形式データを生成する．生成器と識別器の 2 つのニューラルネットワークを同時に学習し，生成器は本物に似た偽データを作り，識別器は本物のデータと偽物のデータを正確に判

[1]名城大学 理工学部 情報工学科
[2]名城大学 大学院理工学研究科 情報工学専攻
[3]藤田医科大学 医学部 医学科

[4]データの使用にあたり，藤田医科大学病院倫理委員会および名城大学倫理審査委員会の承認を得ている．

[5]CTCAE は common terminology criteria for adverse events（有害事象共通用語規準）の略で，有害事象に対する評価基準を指す．CTCAE では有害事象の重症度をグレード 0（正常）〜5（死亡）と定義しており，重症の度合が有害事象ごとに規定されている．

別するように学習を進める．提案者の Xu らは GAN における表形式データのモデル化には，数値などの連続列とカテゴリなどの離散列が混在すること，連続列において複数の分布があること，離散列においてデータ不均衡があることなどの問題を指摘し，その対策として条件付きの生成器や均等で効率的なサンプリングを用いた学習を CTGAN に施している．CTGAN の生成器では元データを直接使用しないため，元データのプライバシー保護には強みがある．

TVAE [11] は，Xu らの実験において CTGAN の性能と比較するために VAE を連続・離散混合型の表形式データに適応させたモデルである．TVAE では元データを直接使用して圧縮，復元を通じて生成器を構築する．CTGAN がランダムなノイズベクトルからデータを生成するのに対し，TVAE は元データを直接使用するため，元データに似たデータを作る点においては CTGAN より有利である．

CTGAN と TVAE は The Synthetic Data Vault (SDV, `https://sdv.dev/`) で提供されているものを利用した．Bourou らはこの SDV の実装を用いて侵入検知システムのデータ生成に関する比較実験を行っている [2].

## 4 実験方法

本研究では加工した訓練データに分類モデルを適用して，その分類精度を測ることで人工データを利用する効果を測る．XGBoost (eXtreme Gradient Boosting) [4] は，勾配ブースティング木と呼ばれるアンサンブル学習モデルである．本研究で用いるデータセットに対して英らは複数の分類モデルについて分類精度を比較し，XGBoost が最も良い分類精度を示したと報告している [12]．そこで本研究では XGBoost を評価用の分類モデルとして使用する．

分類精度の評価においては陽性クラスに対する F 値を主な評価基準とする．ただし，分類モデルの決定閾値を変化させた際に取り得る F 値の中で最大のものを採用する．また，そのときの決定閾値で得られる適合率，再現率，$F_{1.5}$ 値も評価基準として参照する．$F_{1.5}$ 値は下式の $\beta = 1.5$ の場合として計算される．

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

ここで $P$ は適合率，$R$ は再現率である．$\beta$ は再現率に対する重みであるため，F1.5 値は再現率をより重視した指標になっている．通常の F 値は上の $\beta = 1$ の場合に該当する．

Optuna (`https://www.preferred.jp/ja/projects/optuna/`) [1] はハイパーパラメータの最適化を自動化するためのソフトウェアフレームワークである．ハイパーパラメータの試行錯誤を自動的に行いながら，優れたハイパーパラメータの組み合わせを自動的に発見する．本研究では Optuna を XGBoost のハイパーパラメータの決定に使用した．探索アルゴリズムとしてベイズ最適化手法の一種である TPE (tree-structured Parzen estimator) を用いる．Optuna のハイパーパラメータ最適化の目的関数も F 値とする．Optuna 実行時のハイパーパラメータの探索範囲は英ら [12] らと同じにした．実行時間は 24 時間に固定し [6]，疑似乱数によるばらつきを考慮してこの実験を各データセットごとに 3 回ずつ行い，F 値が中央値をとった試行における分類精度の評価値を採用する．

SHAP (SHapley Additive exPlanations) [6] は，協力ゲーム理論における考え方の一つである Shapley 値を機械学習モデルの予測タスクに応用した説明可能 AI 手法である．本研究では，データ拡張等がなされた様々な訓練データで学習した XGBoost が行った各予測に対して各特徴量がどの程度寄与したかを測るために使用する．先述の Bourou らの比較実験 [2] では，生成データの質を比較するのにこのような説明可能 AI 手法までは利用していない．

なお，先述したように，評価実験においては，原データに人工データを加えてデータ拡張したものを訓練データとする場合と人工データのみを訓練データとする場合の 2 通りを試すことにする．

## 5 実験結果

### 5.1 生成データの周辺分布

Raw, CTGAN-10, TVAE-10 データの各々について各特徴量の相対度数分布 [7] を図 1, 図 2, 図 3 に示す．CTGAN-10 データの各特徴量の分布をみると，Raw データにおいてデータ数が少ない値を多く生成していることが分かる．したがって，不均衡なデータにおいて出現頻度の低い値も生成できている．しかし，Raw データの各特徴量の分布を維持できていないことが観察される．一方 TVAE-10 データでは，CTGAN-10 データと比べて Raw データの各特徴量の分布を維持できている．しかし，データ数の少ない値に関して生成されない値がある．Cisplatin dose の値の分布を見ると，Raw データでは 200 [mg/day] 付近の値がわずかに存在するが，TVAE-1 データでは最大値が 150 [mg/day] 付近にあり，出現数の少ない値を生成できていないことが分かる．

### 5.2 分類精度

原データに生成データを加えて分類モデルを学習した場合の分類精度を表 1 に，生成データのみから学習した場合の分類精度を表 2 に示す．どちらの表にも原データのみから学習した場合の分類精度を一番上に記す．

表 1 では，Raw+TVAE-5 では 0.413，Raw+TVAE-20 では 0.404 という F 値が得られた．これは Raw の F 値 0.377 を上回り，陽性クラスの絶対的希少性が緩和された可能性がある．CTGAN では追加する生成データサイズが増加すると精度が下がってしまった．一方，TVAE ではデータサイズと精度の相関はなく，より高い分類精度を出すには追加データサイズを詳細に検討する必要がある．データ拡張の観点で，原データに生成データを加えることで分類精度を上げられることが分かった．

表 2 では，原データがない分，表 1 に比べて全体的に分類精度が下がっている．ただし，TVAE-10 の F 値が 0.396

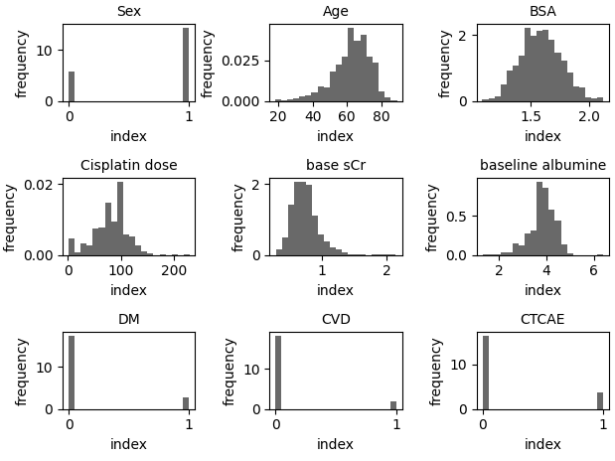



図 1: Raw データにおける各特徴量の分布．

[6] 使用した計算機の CPU は Intel Core i9-9900K である．Optuna はシングルスレッドで実行した．

[7] ビン数を 20 とし，総面積が 1 となるように正規化している．

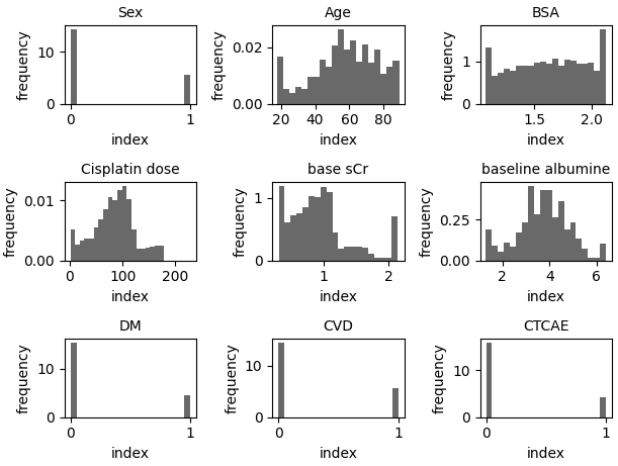



図 2: CTGAN-10 データにおける各特徴量の分布．

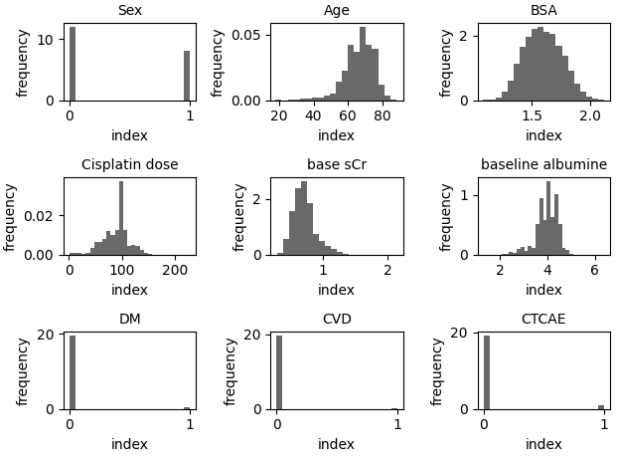



図 3: TVAE-10 データにおける各特徴量の分布．

と最も高くなり，Raw の 0.377 を上回った．やや意外であるが，原データを含まない場合でも原データの分類精度を上回る場合があることが分かった．CTGAN，TVAE どちらの生成データでもデータサイズが増加すると F 値が徐々に上がっていき，データサイズが 20 倍になった時点で下がっている．すなわち，元データの 10 倍から 20 倍のサイズのときに最も高い分類精度が出ている．CTGAN が生成するデータの中では CTGAN-10 の F 値が 0.292 と最も高かったが，Raw より低い値となった．実験全体を通して CTGAN が生成するデータの質は十分とは言えなかった．CTGAN の学習設定が不十分な可能性があるので見直しを行っていきたい．

### 5.3 判断根拠の変化

分類精度が最も高くなった拡張済みデータ Raw+TVAE-5 について SHAP を用いた分類根拠を可視化した．図 4 は各特徴量における SHAP 値の散布図である [8]．各散布図において，横軸がその特徴量の値，縦軸が SHAP 値である．この図における SHAP 値は陽性クラスに対するもので，SHAP 値が大きくなるほど，陽性と判断することへの寄与度が高いことを示す．また，SHAP 値が正であれば陽性と判断することに寄与していること，負の場合は陰性と判断することに寄与

[8]SHAP 値の計算・可視化に利用した Python の shap パッケージでは dependence plot と呼ばれる．

表 1: 原データに生成データを加えた場合の分類精度．

| データセット | F 値 | 適合率 | 再現率 | AUROC | F1.5 値 |
|---|---|---|---|---|---|
| Raw | 0.377 | 0.274 | 0.607 | 0.723 | 0.442 |
| Raw+CTGAN-1 | 0.351 | 0.233 | 0.714 | 0.723 | 0.436 |
| Raw+CTGAN-5 | 0.306 | 0.190 | 0.786 | 0.664 | 0.399 |
| Raw+CTGAN-10 | 0.303 | 0.192 | 0.714 | 0.645 | 0.389 |
| Raw+CTGAN-20 | 0.275 | 0.173 | 0.679 | 0.597 | 0.357 |
| Raw+TVAE-1 | 0.369 | 0.324 | 0.428 | 0.701 | 0.390 |
| Raw+TVAE-5 | 0.413 | 0.371 | 0.464 | 0.719 | 0.431 |
| Raw+TVAE-10 | 0.364 | 0.254 | 0.643 | 0.715 | 0.437 |
| Raw+TVAE-20 | 0.404 | 0.288 | 0.679 | 0.721 | 0.479 |

表 2: 生成データのみから学習した場合の分類精度．

| データセット | F 値 | 適合率 | 再現率 | AUROC | F1.5 値 |
|---|---|---|---|---|---|
| Raw | 0.377 | 0.274 | 0.607 | 0.723 | 0.442 |
| CTGAN-1 | 0.274 | 0.177 | 0.607 | 0.550 | 0.347 |
| CTGAN-5 | 0.278 | 0.172 | 0.714 | 0.605 | 0.363 |
| CTGAN-10 | 0.292 | 0.183 | 0.714 | 0.593 | 0.378 |
| CTGAN-20 | 0.266 | 0.156 | 0.929 | 0.578 | 0.367 |
| TVAE-1 | 0.315 | 0.230 | 0.500 | 0.624 | 0.367 |
| TVAE-5 | 0.333 | 0.235 | 0.571 | 0.664 | 0.397 |
| TVAE-10 | 0.396 | 0.274 | 0.714 | 0.713 | 0.478 |
| TVAE-20 | 0.362 | 0.258 | 0.607 | 0.707 | 0.423 |

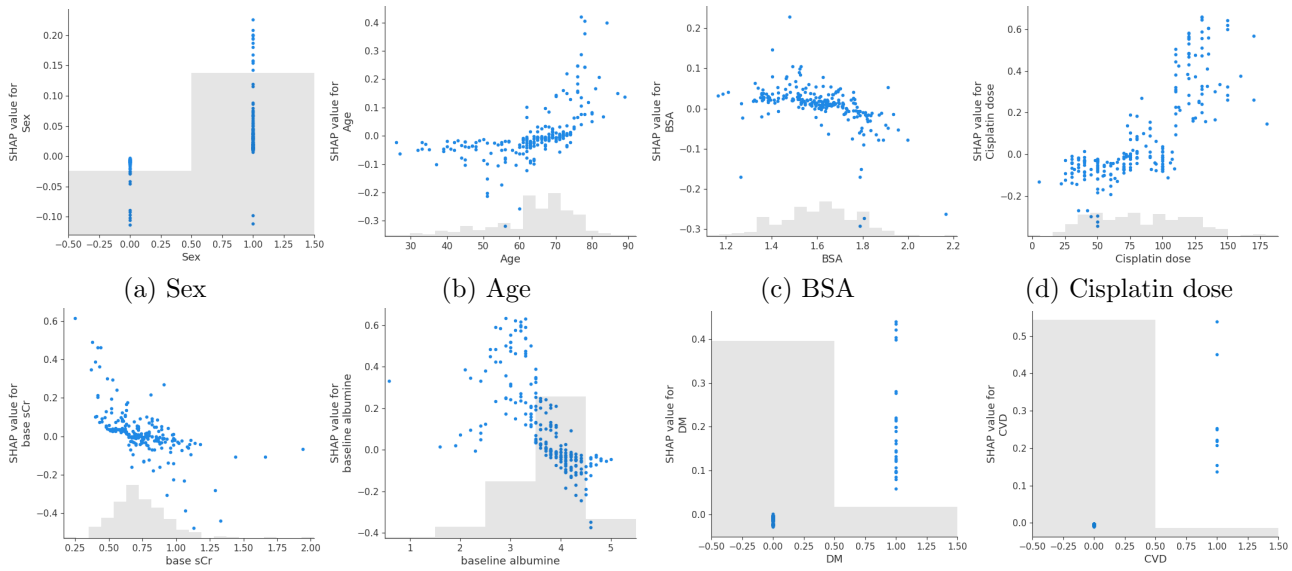
していること，0 の場合は陽性・陰性いずれにも寄与がないことを意味する．青い点の一つ一つがテストデータ中の患者事例に対応する．背景に描画されている薄い灰色のヒストグラムはテストデータの分布を表している．

Cisplatin dose，base sCr，baseline albumine の散布図を観察すると，図 4 (d) より，Cisplatin dose の値が 75∼100 [mg/day] を境に値が大きくなると SHAP 値が高くなっている．また，図 4 (e) より base sCr の値が 0.75 [mg/dl] を境に値が小さくなると SHAP 値が高くなる．同様に，図 4 (f) より baseline albumine の値が 4.0 [g/dl] を境に値が小さくなると SHAP 値が高くなる．これらの傾向は，英らの実験における原データでの実験結果と同じである [12]．従って，原データを TVAE-5 で拡張したデータから学習した分類器の分類根拠は原データのみから学習した分類モデルのものと整合性がとれているといえる．

## 6 おわりに

表形式のデータに基づく分類モデルの学習において深層生成モデルを用いたデータ拡張を行うことを提案し，その有効性を評価した．そして評価実験の結果，深層生成モデルを用いたデータ拡張により分類精度を向上させることが可能であるという知見を得た．また，CTGAN が生成するデータに比べて TVAE が生成するデータの方が高い分類精度を出すことが分かった．

一方で，CTGAN の生成器では原データを直接使用しないため，プライバシー保護の観点では CTGAN の方が優れていると言える．原データを用いずに人工データのみを利用する場合には（特に TVAE に対して）プライバシー保護の観点からの評価も必要であると思われる．また，その他の課題として，SMOTE (Synthetic Minority Over-sampling Technique) [3] のようなオーバーサンプリング手法との比較，別のデータへの適用などが挙げられる．




(a) Sex (b) Age (c) BSA (d) Cisplatin dose

(e) base sCr (f) baseline albumine (g) DM (h) CVD

図 4: TVAE-5 データを用いて構築した XGBoost における SHAP 値の散布図.

## 参考文献


[1] Akiba, T., et al.: Optuna: A next-generation hyperparameter optimization framework. In: Proc. of KDD-19 (2019)

[2] Bourou, S., et al.: A review of tabular data synthesis using GANs on an IDS dataset. Information 12(9) (2021)

[3] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. J. of Artificial Intelligence Research 16 (2002)

[4] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. of KDD-16 (2016)

[5] Han, C., Murao, K., Satoh, S., Nakayama, H.: Learning more with less: GAN-based medical image augmentation. Medical Imaging Technology 37(3) (2019)

[6] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proc. of NIPS-17 (2017)

[7] Okawa, T., et al.: Prediction model of acute kidney injury induced by cisplatin in older adults using a machine learning algorithm. PLOS ONE 17(1) (2022)

[8] Park, N., et al.: Data synthesis based on generative adversarial networks. In: Proc. of VLDB-18 (2018)

[9] Sorin, V., Barash, Y., Konen, E., Klang, E.: Creating artificial images for radiology applications using generative adversarial networks (GANs) — a systematic review. Academic Radiology 27(8) (2020)

[10] Weiss, G.M.: Foundations of imbalanced learning. In: He, H., Ma, Y. (eds.) Imbalanced Learning. Wiley-IEEE Press (2013)

[11] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Proc. of NeurIPS-19 (2019)

[12] 英彰悟, 亀谷由隆, 水野智博: シスプラチン誘発性急性腎障害の発症を予測する機械学習モデルの構築と予測根拠の分析. In: 第 12 回日本医療情報学会「医用人工知能研究会」・人工知能学会「医用人工知能研究会」合同研究会予稿集. SIG-AIMED-012-04 (2022)

[13] 高橋克己: プライバシー保護データマイニング. システム/制御/情報 63(2) (2019)