URL https://journal.ntt.co.jp/article/23456
DOI https://doi.org/10.60249/23105102

パーソナルデータ利活用促進に向けた匿名化・合成データ生成技術にかかわる取り組み

本稿では，世の中の匿名加工技術・合成データ生成技術の実用化動向を紹介するとともに，NTTテクノクロスの匿名加工情報作成ソフトウェア「tasokarena：タソカレナ」で適用している匿名加工技術・合成データ生成技術および群馬大学と連携した共同研究（合成データ生成技術に関する最適な実装方式の共同研究）の取り組み内容を紹介します．

キーワード：#匿名加工情報，#合成データ，#個人情報

石原 一郎[†1] / 角田 進[†1]
宮木 一郎[†1] / 吉岡 甲将[†1]
千田 浩司[†2]

NTTテクノクロス[†1]
群馬大学[†2]

はじめに

昨今，国内DX（デジタルトランスフォーメーション）が加速し，データ利活用が注目されています．その中で，個人情報から特定の個人を識別できないように加工した匿名加工情報は，個人情報保護法や次世代医療基盤法などの関連する法整備とともに利活用が拡大しています．政府機関の個人情報保護委員会によると，匿名加工情報の作成・提供に関する公表を行っている事業者が2020年３月時点で500社にのぼっています[*1]．

海外では個人情報の取り扱いの法令としてGDPR（General Data Protection Regulation：EU一般データ保護規則）やCCPA（California Consumer Privacy Act：カリフォルニア州消費者プライバシー法）がありますが，日本の「匿名加工情報」のようなデータの定義・利用範囲や具体的なデータ加工方針については言及されておらず，多人数のデータの統計的特徴に基づき人工的に作成される架空のデータである合成データ[*2]の利用がさかんです．

匿名加工技術および合成データ生成技術の実用化動向

一般に匿名化というと，誰の情報か分からなくする，氏名を取り除くといったイメージがあるかもしれません．個人情報保護委員会が公開しているFAQ[*3]によれば，匿名化は個人情報から氏名，生年月日，住所，個人識別符号[*4]等，個人を識別することができる情報を取り除くこと，ただし匿名化を行ってもなお特定の個人が識別できる場合には個人情報に該当する，とあります．そして匿名加工情報は，個人情報を個人情報保護委員会規則で定める基準（個人情報の保護に関する法律施行規則[*5]第34条第１～５号）に従って加工したものであり，当該個人情報を復元して特定の個人を再識別できないようにする必要があります（個人情報保護法第２条第６項）．なお本稿では特定の個人に関するデータ（レコード）の集合をパーソナルデータと呼び，１人以上の個人に関するデータであれば，個人情報も匿名加工情報もパーソナルデータに含まれるものとします．

それでは，匿名加工情報を作成するための匿名加工技術にはどのようなものがあるでしょうか．個人情報保護委員会のガイドライン[*6]では個人情報の保護に関する法律施行規則第34条第１～５号を満たす具体例とともに，具体的な匿名加工技術の手法が例示されています（**表**）．そして実際には，匿名性と有用性のバランスの取れた適切な加工が重要となります．ここで有用性とは，所望の利活用において匿名加工情報が元の個人情報と比べてどの程度利用価値を維持しているかを示す指標です．加工が不適切だと，匿名性が損なわれた情報，有用性の低い匿名加工情報となってしまいます．

もっとも有名な匿名性に関する指標の１つとして，k-匿名性[(1)]があります．性別や年齢のように特定の個人を絞り込める属性（準識別子）の値の組について，k人以上が同じ値になるとき，そのパーソナルデータはk-匿名性を満たすといいます．k-匿名性を満たすパーソナルデータを作成するために，表の手法などが用いられます．また近年では，差分プライバシー[(2)]と呼ばれる匿名性・プライバシーに関する指標の研究も進展しています．差分プライバシーは，パーソナルデータに限らず，パーソナルデータから得られる統計値や機械学習・深層学習の生成モデルにも適用できます．統計値や生成モデルからも特定の個人に関するデータを推定されるリスクがあり，攻撃も年々高度化していることから，差分プライバシーの注目度も高まっています．なお匿名加工情報は必ずしもk-匿名性や差分プライバシーを満たす必要はありませんが，これらを満たすことにより匿名性・安全性を理論的に保証できる効果があります．

＊1 パーソナルデータの適正な利活用の在り方に関する実態調査（令和元年度）報告書（個人情報保護委員会）：https://www.ppc.go.jp/files/pdf/personal_date_report2019_1.pdf
＊2 合成データ：個人情報保護委員会のガイドライン[*6]の別表2では疑似データと表現していますが，本稿では合成データと統一します．
＊3 「匿名化」された情報と「匿名加工情報」との違い：https://www.ppc.go.jp/all_faq_index/faq3-q2-12/
＊4 個人識別符号：特定の個人を識別することができるものとして政令に定められた文字，番号，記号その他の符号．指紋や静脈などの身体的特徴を表した符号，運転免許証の番号やマイナンバーなど．
＊5 個人情報の保護に関する法律施行規則：https://elaws.e-gov.go.jp/document?lawid=428M60020000003
＊6 個人情報の保護に関する法律についてのガイドライン（仮名加工情報・匿名加工情報編）：https://www.ppc.go.jp/personalinfo/legal/guidelines_anonymous/

表　匿名加工技術の例（個人情報保護委員会のガイドライン*6の別表2を引用）

手法名	解説
項目削除／レコード削除／セル削除	加工対象となる個人情報データベース等に含まれる個人情報の記述等を削除するもの．例えば，年齢のデータを全ての個人情報から削除すること（項目削除），特定の個人の情報を全て削除すること（レコード削除），又は特定の個人の年齢のデータを削除すること（セル削除）．
一般化	加工対象となる情報に含まれる記述等について，上位概念若しくは数値に置き換えること又は数値を四捨五入などして丸めることとするもの．例えば，購買履歴のデータで「きゅうり」を「野菜」に置き換えること．
トップ（ボトム）コーディング	加工対象となる個人情報データベース等に含まれる数値に対して，特に大きい又は小さい数値をまとめることとするもの．例えば，年齢に関するデータで，80歳以上の数値データを「80歳以上」というデータにまとめること．
ミクロアグリゲーション	加工対象となる個人情報データベース等を構成する個人情報をグループ化した後，グループの代表的な記述等に置き換えることとするもの．
データ交換（スワップ）	加工対象となる個人情報データベース等を構成する個人情報相互に含まれる記述等を（確率的に）入れ替えることとするもの．
ノイズ（誤差）付加	一定の分布に従った乱数的な数値を付加することにより，他の任意の数値へと置き換えることとするもの．
疑似データ生成	人工的な合成データを作成し，これを加工対象となる個人情報データベース等に含ませることとするもの．

一方，機械学習・深層学習を用いた合成データ生成技術の研究開発および実用化も急速に進展しており，特に画像データは非常に高品質の合成データが作成可能となっています．構造化された表形式のパーソナルデータについても有用性に優れた合成データ生成技術がいくつか提案されており，諸外国では多くのスタートアップ企業がパーソナルデータを対象とした合成データ生成事業を行っています[3]．興味深い話として，自社の作成する合成データがGDPRやCCPAの匿名性要件に準拠していると主張している企業もみられます．しかし筆者らが知る限り，国内では匿名加工情報の基準を満たす，あるいは非個人情報と認められるような合成データの要件に関する議論はほとんど行われていません．

上記を踏まえ，安心・安全なパーソナルデータ利活用促進に向け，データ合成技術評価委員会が国内で発足しました*7．合成データを構成する各レコードは架空のデータであり，一般に特定の個人とは紐付きません．また合成データは多属性のパーソナルデータでも属性間の統計的特徴を維持しやすく，その有用性の高さが注目されています．しかし特定の個人に関するデータの推定リスクは，合成データ生成技術の手法に依存します．そこでデータ合成技術評価委員会では「不適切な合成データの利用」や「リスクを恐れた合成データの利用躊躇」の課題を解決し，健全な合成データ生成技術の利用を推進するため，既存の合成データ生成技術の匿名性やプライバシーレベルを評価し，結果を発信していくことをめざしています．筆者らもデータ合成技術評価委員会の活動に参画しています．差分プライバシーを満たす合成データ生成技術も多数提案されており，前述のスタートアップ企業の一部がすでに実用化していることから，国内でも健全な合成データ生成技術の実用化が進み，データ利活用による社会課題の解決や安心・安全で便利なサービスの普及に資することが期待されます．

tasokarenaで適用している技術と特長

NTTテクノクロスでは「匿名加工情報」の作成を支援するソフトウェア（tasokarena：タソカレナ）[4]を2018年から提供開始し，2021年から合成データの生成機能を追加実装しています．現在では医療・金融・自治体・コールセンタ等さまざまな分野へ導入されています．tasokarenaの製品名は，日が暮れて薄暗くなり相手の顔の見分けがつきにくくなったころに「あなたは誰ですか？」と問いかける言葉「誰そ彼（たそかれ）」が，元のデータから個人を特定できなくする本ソフトウェアのコンセプトと合致していることから命名しています．

tasokarenaの主な特長を紹介します．

⑴　NTT独自技術含む豊富な加工技法を提供

数十種類の加工技法の中から実行する加工技法を組み合わせ，匿名加工情報を作成することが可能です．

特徴的な加工技法としてはNTTが独自に開発した手法である「Pk-匿名化[5]」を実用化し，本ソフトウェアに搭載しています．匿名性の代表的な指標であるk-匿名性を満たすようにノイズ（疑似データ）の付与やデータの入れ替えを行い，データの有用性が損なわれていない匿名加工情報を作成します（**図**）．

また，受診履歴データや購買履歴データといった履歴型データ（1ユーザ複数レコード）についても，k-匿名化，Pk-匿名化を行うアルゴリズムを実装し，履歴型データについても加工と評価を実行することも可能です．

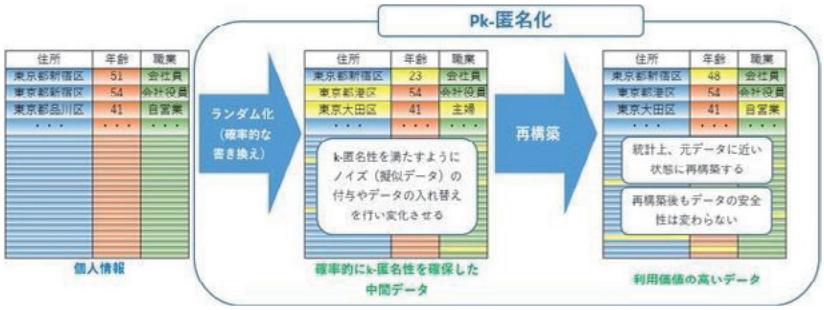
⑵　匿名性・有用性の評価機能

tasokarenaで作成した匿名加工情報は，15種類の評価技法により匿名性と有用性のバランスをグラフで確認することが可能です．このグラフを参考にしながら加工技法の組合せを変えることで，最適な加工ルールを設定していくことが可能になります．

⑶　マスキングツールで情報共有の安全性を向上

同一事業者内での情報共有の安全性を高める目的として，自由記述形式で記載された文章に含まれる個人情報を削除するマスキングツールを提供しています．医師の所

＊7　データ合成技術評価委員会：https://www.iwsec.org/pws/ppsd



NTT公式HPから引用
https://www.rd.ntt/research/PF99-341.html

図　Pk 匿名化イメージ

見や患者の診療記録，コールセンタの対応履歴データ等の文章に含まれる個人情報に対して，自然言語解析によって氏名や住所などを自動で判別し，削除することが可能です．これにより，ユーザの作業負担を軽減しながら，情報共有における安全性の向上を実現しています．

⑷　医療向けオプション

自治体・医療機関・健康保険組合などの共通仕様になっているレセプト*8データをtasokarenaで読み込み可能にする変換ツールを提供しています．レセプトデータのフォーマットはレコード識別情報の値によってレコード項目数や記録内容の形式が異なるため，匿名加工情報を作成する場合には，事前にユーザによるデータの形式を合わせるなどのクレンジング処理が必要でした．変換ツールを使用することでユーザによる手間が削減され，さらに変換後のレセプトデータから作成した匿名加工情報を再び元のレセプトデータのフォーマットに戻すことが可能になり，既存システムでも匿名加工情報を利用できます．

また，医療系のデータを扱う際の標準の1つであるPhUSEの非特定化標準*9を基に，匿名加工したデータセットの整合性をとるツールを提供しています．

⑸　合成データ生成機能

合成データ生成機能として統計的手法と機械学習的手法を提供しています．

統計的手法はNTT社会情報研究所の特許技術を活用しています．各属性の平均など統計値が元データとほぼ等しい合成データを生成する技術等を独自に開発し，これまでプライバシ保護技術では実現できなかった分析に必要な複数の統計値を保持する多属性の合成データを生成することが可能になりました．NTT社会情報研究所は本技術の開発で培った知見を活用し，AI（人工知能）・機械学習分野における難関国際会議の匿名化技術を競うコンペティションで優勝しました(6)．

機械学習的手法はベイジアンネットワーク*10を基に合成データを生成します．

NTTテクノクロスと群馬大学との共同研究の取り組み

合成データについては具体的な利用例はありますが，安全性，有用性についての評価方法は定まっていません．また，合成データ生成手法としてさまざまな技術が提案されていますがどのようなデータにどのような手法を用いると良いかといった確立されたノウハウやコンセンサスがあるわけではありません．そこで，NTTテクノクロスと群馬大学で市中の合成データ生成手法による合成データ生成，および安全性，有用性評価を行い，データ種別による各合成データ生成手法の得手不得手を明らかにする共同研究を行っています．

今後の展開

NTTテクノクロスは，今後，共同研究の取り組みを継続実施していくとともに，その結果を，NTTテクノクロスの製品へ組み込み，ビジネス展開することをめざしています．また，個人情報保護法は個人情報の保護に関する国際的，技術状況等を勘案し，3年ごとに必要に応じて改正されることになっており，個人情報保護法の動向をチェックし，適切な技術を提供することを通じてパーソナルデータの利活用の拡大に貢献していきます．

■参考文献

(1) L. Sweeney：“k-anonymity: A model for protecting privacy,” Int. J. Uncertain. Fuzziness Knowl.-Based Syst., Vol. 10, No. 5, pp. 557–570, 2002.
(2) C. Dwork：“Differential privacy,” Proc. of 33rd ICALP – Volume Part II, LNTCS, Vol. 4052, pp. 1–12, 2006.
(3) 千田・南・寺田・伊藤：“プライバシー保護型合成データの実用動向と今後の展望，”統計，Vol. 73, No. 8, 2022.
(4) https://www.ntt-tx.co.jp/products/anontool/
(5) https://www.rd.ntt/research/PF99-341.html
(6) https://group.ntt/jp/newsrelease/2021/03/02/210302b.html


（左から）　宮木　一郎／石原　一郎／
角田　　進／吉岡　甲将／
千田　浩司

個人情報をより安全に守りながら，より効果的に活用し，企業や個人に利益をもたらすことをめざし取り組みを進めています．

◆問い合わせ先

NTTテクノクロス
セキュアシステム事業部　tasokarena担当
TEL　045-212-7577
E-mail　anontool.info-ml@ntt-tx.co.jp

＊8 レセプト（診療報酬明細書）：医療費の請求明細のことで，保険医療機関・保険薬局が保険者に医療費を請求する際に使用するものです．電子レセプトとは，厚生労働省が定めた規格・方式（記録条件仕様）に基づきレセプト電算処理マスターコードを使って，CSV形式のテキストで電子的に記録されたレセプトのことを指します．

＊9 Pharmaceutical Users Software Exchange：https://phuse.global/

＊10 ベイジアンネットワーク：データの因果関係の強さを，ある事象が起こった場合に他の事象が起こる確率の大きさから判断し，多数の事象間の因果関係をグラフィカルに整理する方法．