The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

残差に基づいて匿名性と有用性を両立させる匿名加工に関する考察

A Consideration on Anonymization of Personal Information to Establish Both Anonymity and Usability Based on Residuals

森下 壮一郎 *1
Soichiro Morishita

*1株式会社サイバーエージェント
CyberAgent, Inc.

In Japan, the revised Personal Information Protection Act is enacted in full force on May 30, 2016. It became possible to provide personal information to third parties under certain conditions such as anonymization without user agreements. However, in the case of no limitation on background knowledge of attackers, complete anonymization maintaining usability of data is impossible in principle. It is difficult to establish a rational anonymous processing standard. In this paper, an anonymization method based on residuals to establish both anonymity and usability is proposed. Moreover, it is shown that to apply anonymization with ambiguous purpose of use is meaningless.

1. はじめに

2016 年 5 月 30 日に全面施行された改正個人情報保護法（以下，法）において，匿名加工情報については一定の条件の下で本人の同意なしで第三者提供できるようになった．しかしながら個人情報保護委員会規則で定める基準は包括条項を含んでおり要件が不明瞭であることから，個人情報取扱事業者による匿名加工情報としての第三者提供および利用は進んでおらず，多くの場合で,「共同利用」や「業務委託」での利用になっている．これらも同意なしで利用可能な様態であるが，広くデータを利活用するという観点からは望ましくない．

一方で，攻撃者が持つ背景知識に限定がない場合，データの有用性を保ちながらの完全な匿名化は原理的に不可能であるので，合理的な匿名加工の基準を定めることは困難である．この難点を解消するには，背景知識を用いた個人の特定や情報の詳細化について法制などで制約を課すか，データの有用性について譲歩するかのいずれかの条件を満たす必要がある．前者のコントロールには限界があり，またリスク評価が困難である．一方後者は，実質的にデータの意味がなくなるという副作用は予想されるものの，データの利活用の目的を限定すれば，その条件下での最低限の効用は損なわない最大限の匿名化を施すことができる．以上の考えの下，本稿ではデータの有用性を保ちながらの個人情報の匿名加工の実施手順について検討する．

2. 匿名加工の基準および利用に関する指針

匿名加工の方法については「個人情報保護委員会規則」（以下，規則）で定める基準に従う必要がある〈法 36 条 1 項〉．しかしながらガイドラインでは例示があるのみで具体的な手法については確立されていない．法第 36 条第 1 項の個人情報保護委員会規則で定める基準（以下，基準）では，匿名加工で講ずるべき措置について 1 号から 5 号まで挙げている．国立情報学研究所の匿名加工情報に関する技術検討ワーキンググループは，基準の各号で示された措置について考察して「匿名加工情報の適正な加工の方法に関する報告書」としてまとめている [国立情報学研究所 17]．この報告書で，1 号から 4 号の措置については，曖昧性を排除して要件を定義した上で厳格な適用を求めている．5 号は包括条項であるので，この措置の要件については，匿名加工情報の有用性を損なわないことを重視して無限定の背景知識は仮定せず，特定対象項目（1 号の措置の対象で，特定の個人を識別することができる項目）の容易照合性について「通常の業務における一般的な方法で可能な状態」としている．そして，情報を結合する行為について特定対象項目の情報を増加や詳細化するようなものを識別行為禁止義務〈法 36 条 5 項〉違反と解釈している．

これは，前述の 2 つの条件のうちの「背景知識を用いた個人の特定や情報の詳細化について何らかの限定を施す」ことで合理的な匿名加工の基準を定めるときの難点を解消しようとするものである．これにより匿名加工の実施手順については明確になった．しかしながら識別行為禁止義務違反とならないような利用の形態については曖昧性が残っていた．

情報法制研究所のオンライン広告研究タスクフォースは,「オーディエンスターゲティング広告における匿名加工情報の利用に関する提言」（以下，提言）でオンライン広告における匿名加工情報の活用方法に関する指針を示した [情報法制研究所 17]．この提言では匿名加工の方法については議論していない．一方，報告書等で「識別行為禁止義務」違反とされている行為についての具体的な検討を行っており，適法となる形態（3.2.4 項）と違法となる形態（2.2.3 項）とをそれぞれ示している．

これらの報告書と提言，および他のプライバシー保護技術関連の教科書やガイドラインなどの匿名加工に関する文書では，プライバシーとデータの有用性とは常にトレードオフの関係にあると説明される．本稿ではこのトレードオフについて，報告書および提言の用語を用いながら検討する．

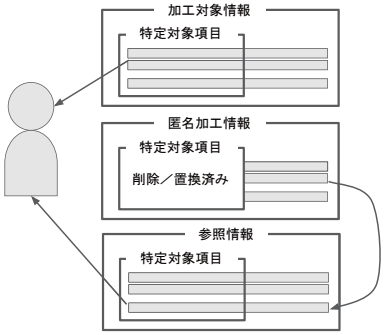
3. トレードオフについての検討

報告書では，匿名加工の対象となる個人情報を「加工対象情報」，基準の 1 号の措置の対象になる項目を「特定対象項目」，加工対象情報の外部にあって，それを参照することにより特定の個人が識別できる情報を「参照情報」と呼んでいる．これらの関係を図 1 に示した．

匿名加工情報には特定対象項目が含まれていないにも関わらず，参照情報に共通に含まれる項目によって，参照情報の特定対象項目と紐付けられて個人の特定に至っている．このような働きをする項目は，一般に準識別子と呼ばれる．無限定の背

連絡先: 森下 壮一郎，株式会社サイバーエージェント 秋葉原ラボ，morishita_soichiro@cyberagent.co.jp

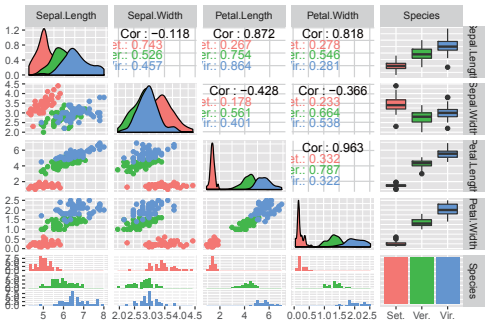
図 1: 参照情報による個人の識別



景知識（参照情報）を前提とするとき，当たり障りがないような項目からでも個人の特定に至る危険がある．

個人の特定に至らないまでも情報が詳細化される場合もある．わかりやすさのために，判別分析のテストデータとしてよく用いられる Iris データセット [Fisher 36] を仮想的に個人に関する情報と見なして詳細を述べる．図 2 にデータの分布を示す．作図には統計ソフトの R[R Core Team 17] を用いた．

図 2: Iris データセットにおけるデータの分布



これはアヤメの品種（species）と，がく（sepal）および花弁（petal）の長さと幅をそれぞれ計測したものである．品種ごと，すなわちクラスごとに各変量の平均が有意に異なり，また変量間にある程度の相関がある．個人に関する情報がこのように分布することは多い（例えば，出身学校と試験の各科目の成績など）．したがって，これを仮想的に個人に関する情報とみなすことに妥当性がある．なお，ここでは各品種における各変量の分布を攻撃者が背景知識の一部として持っているとする．

このデータセットから仮に品種の項目を削除したとしても，既知の分布から判別分析により品種を推定することは容易である．これは，背景知識に基づいて情報を詳細化することに対応する．統計情報として平均や分散が公知であることは多いので，この対応付けは妥当である．逆に，品種の項目を残して花弁の長さと幅の項目を削除したとしても，この 2 つの変量はクラス間分散に対してクラス内分散が小さいので，統計情報を基に一定の確かさで情報を詳細化できる．クラス間分散に対してクラス内分散が充分に大きくなる加工を施せば詳細化のリスクは小さくなるが，同時にクラスの違いが持つ情報量が小さくなる．これが匿名性と有用性のトレードオフの要因である．

4. 残差に基づく匿名加工

本稿では，以上で述べた状況下で有効な匿名加工の手続きとして，残差に基づく匿名加工を示す．例えば前述の Iris データにおいて，各クラスのクラス内分散共分散に基づいて正規化を行って，平均との差を残差として改めて特徴量とする．この操作によりすべての特徴量で平均が 0，分散が 1 になるので，詳細化のリスクはほとんどなくなる一方で，有用性は完全に失われるように見える．しかしながら実は目的によっては各クラスにおける残差が意味のある特徴量であることは多い．Iris データにおいては，品種ごとに平均が異なるところを補正したと考えることができる．すなわち，「この品種の中では大きい／小さい」という相対的な評価軸として変量を改めて定義したことになる（個人に関する情報でも，例えば各出身学校における相対的な成績の方にむしろ意味があるなどの場合がある）．なお削除対象の項目による分布の偏りを正規化できれば良いので，削除対象の項目が連続変数である場合には，例えば一般化線形モデルによる回帰分析における残差を用いることで同様の効果を得ることができる．

5. 議論

本項で述べた情報の詳細化の考え方に基づくと，元になる情報が法令上の個人情報でない（特定対象項目を含まない）場合でも，個人に関する情報であれば参照情報により個人が特定されたり情報が詳細化されたりするリスクがある．いずれにしても本稿で示した手順により，情報が詳細化されるリスクは軽減できるが，残差が有用な特徴量となりうるかは利用目的に応じて判断する必要がある．これはすなわち，「利用目的が曖昧なままに匿名加工を施すことにそもそもの意味がない」ことに他ならない．原則的には詳細化等のリスクについては識別行為禁止義務違反で処理すべきであるが，無目的なデータ提供における有用性を盾にした不充分な匿名化の妥当性は問われるべきであろう．なお例示の中に，統計量が参照情報となり情報が詳細化されるものがあった．そのような統計量を公知のものとしてよいのかという疑問も浮かぶが，本稿で示した手順に則ればその限りは問題ないことになる．これは，統計量への集計の入力は個人情報の利用としない現行法制と平仄が合っている．

6. おわりに

本稿では，攻撃者の背景知識を限定できない状況下での情報の詳細化のリスクと，それを充分に軽減しながらも有用性を保つ手法として残差に基づく匿名加工について述べた．今後は実施手順の詳細化を進める．

参考文献

[Fisher 36] Fisher, R. A.: The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, Vol. 7, No. 2, pp. 179–188 (1936)

[R Core Team 17] R Core Team, : *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2017)

[国立情報学研究所 17] 国立情報学研究所 匿名加工情報に関する技術検討ワーキンググループ:匿名加工情報の適正な加工の方法に関する報告書, `http://www.nii.ac.jp/research/reports/pd/report-kihon-20170221.pdf` (2017)

[情報法制研究所 17] 情報法制研究所 オンライン広告研究タスクフォース：オーディエンスターゲティング広告における匿名加工情報の利用に関する提言, `https://www.jilis.org/pub/pub_priv20171216.pdf` (2017)