2M4-OS-10a-6

# パーソナルデータ二次利用の有効性向上を目的とした、統計的匿名加工手法の研究

## Research for statistical anonymity processing techniques for effective improvement of the personal data secondary use

泉 晃[*1]
Akira Izumi

尾崎 幸謙[*2]
Koken Ozaki

[*1*2] 筑波大学大学院ビジネス科学研究科
Graduate School of Business Sciences, University of Tsukuba #1#2

When utilizing secondary benefit due to circulation of the personal data, it is necessary to the use and analysis while achieving privacy protection. But as a method for the k-anonymity and the like it has been proposed, in order to ensure an individual's specific resistance by suppressing or generalize the information, and the decline of usefulness due to information loss is a problem in the context of use. This Research use statistical method in order to prevent the usefulness of the deterioration of the utilization.

## 1. 概要

個人情報を含むパーソナルデータの流通による二次的な利活用する際には、プライバシー保護を実現しながら活用・分析をすることが必要となる。そのための手法として k-匿名化等が提唱されているが、情報を抑圧もしくは一般化することにより個人の特定性を担保するため、活用・分析の場面では情報損失による有用性の低下が問題となる。そこで、本研究では活用・分析の有用性の劣化を防止するために統計的手法の適用を研究したい。

## 2. テーマ選定に至った背景

近年、ビッグデータの活用については様々な領域に置いて積極的な取り組みが見られる。なかでも 2011 年 11 月に世界経済フォーラムで報告された「パーソナルデータ:新たな資産カテゴリーの出現」において、「パーソナルデータは、インターネットにおける新しい石油であり、デジタル世界における新たな通貨である」[1]と表現された様に、パーソナルデータの活用については今後更なる発展が期待されている。

また、今後のパーソナルデータの利活用の促進に向けて、2015 には個人情報保護法の初の大規模な改正され成立した。しかしながら、駅の乗降履歴データの販売停止に代表されるように、個人のプライバシーを保護した上でどの様にパーソナルデータを利活用するかについては課題となっている。

パーソナルデータの匿名化については、今後分析や事業利用が広がっていく事を鑑みると、第三者へ提供する場合のみならず企業内やグループ企業内での利用する場合でもあっても、生の個人情報やパーソナルデータは隔離されたセキュアな環境で限定された管理者のみがアクセス可能とし、匿名化されたデータを分析・利用する方式を取るなど、セキュリティ上の要請により匿名化と、そのための匿名化の具体的手法の研究開発が求められる。

## 3. 問題設定

個人情報およびパーソナルデータについて、プライバシー保護を実現しながら分析活用をする為の技術として、プライバシー保護データマイニングと呼ばれる PPDM（privacy-preserving data mining）の研究が近年されている。

PPDM には大きく分けて 3 つの方向性がある。

（1）データそのものを秘匿し分散させた上で演算処理関数を行うマルチパーティ計算方式（暗号学的・MPC アプローチ）[2]

（2）出力されたデータに対してノイズを加える摂動方式[2]

（3）データに対して、一般化・抑圧（削除）を行う匿名化方式[3]である。

現在のところ PPDM を実務的に使用されている例としては多くないが、データの匿名化自体は様々な領域で行われている。今後の PPDM の本格的な発展を考えると複数の方式を組み合わせて使用することも想定されるが、個人の特定性を低減と言う観点からは匿名化方式における手法開発が重要であると思われる。

匿名化方式について現状を概観すると、静的なデータに対しての匿名化技術として、k-匿名化と、その派生である l-多様性と、さらに k-匿名化を確率空間へと拡張させた P k-匿名化と呼ばれる方式などがあり、逐次データに対する匿名化方式として、m-不変性、l-希少性 、多重 P k-匿名化などがある。

k-匿名化とはデータから一個人の識別が出来ない様に同じデータの組みあわせが少なくとも k 個以上存在する様にする方式である。具体的には、名前・ID などの識別子となる情報は削除し、組み合わせると個人の識別が可能となる年齢・住所・行動履歴などの準識別子を一般化・抽象化する。

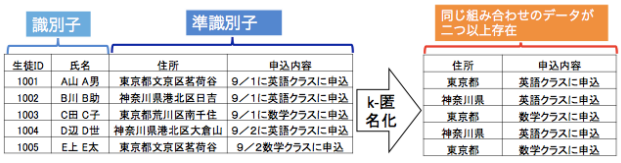


図 1:k-匿名化例

一般に、識別子の削除や仮名化だけでは匿名化にはならないと考えられる。なぜなら、属性の組合せで個人を特定する事が可能となるからである。

実際に匿名化が問題となった事例として、2008 年に米国のオンライン DVD レンタルサービスのレコメンドアルゴリズムコンテストでは外部データソースと突合した Linked Attack により個人特定をされた事例がある。本事例ではデータが匿名化されていたにも関わらず、当該サービスである作品を借りた人が、同じ時期に別サービスで同タイトルについてのレビューを書いていたため、個人の特定に結びつき、それが機微情報にあたる内容で

あったためプライバシー訴訟にまで発展している(後に和解、当コンテストは FTC の指摘を受け 2009 年で終わっている)。

以上の様に、パーソナルデータの利用・分析を行うためには k-匿名化などの具体的な加工方式の検討が必要となると考える。

## 4. 従来手法の課題

従来の匿名化の研究は、主にどの様に匿名化をすればプライバシーが保護されるかという面を重点に研究が行われてきた。一方でパーソナルデータの活用や分析の面を考えると、匿名化を行う事は情報の精度を落とす事につながり、また盲目的に匿名化を行ってしまうと必要なデータが削除され活用が出来ないリスクも考えられる。

実際に「Suica データ社外提供是非に関する有識者会議」では k-匿名化の評価を行っているが、その中では「匿名化した Suica 分析用データを作成したが、k- 匿名化処理によって、相当数の利用者のデータが使用できなくなり、分析結果の精度が低下したり、分析の種類が制限されうる課題があることがわかった」と述べられている[4]。

k-匿名化には様々なアルゴリズムが提唱されているが一般的には下記の手法などが利用される。

・Incognite[5]:フルドメイン一般化
・Mondrian[6]:ドメイン空間の領域分割
・クラスタリングによる k-匿名化[7]

実際にデータに対して Incognite アルゴリズムによる k=2 の匿名化を行った。利用するデータは Mondrian アルゴリズムのデモデータを利用した、内容としては成人属性と資産額についての 98 人分のデータであり 5 つの準識別子と 1 つの機密属性を持つ。デモデータの項目と値は以下の通りである。

| | |
|---|---|
| Age(7値): | 10, 20, 30, 40, 50, 60, 70 |
| WorkClass(6値): | Federal-gov, State-gov, Local-gov, private, Self-emp-inc, Self-emp-not-inc |
| Education(12値): | 5th-6th, 7th-8th, 9th, 11th, Assoc-voc, Assoc-acdm, HS-grad, Prof-school, Some-collge, Bachilors, Masters, Doctorate |
| MatritalStatus(4値): | Married-civ-spouse, Divorced, Separated, Never-married |
| Occupation(12値): | Adm-clerical, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspect, Other-service, Sales, Prof-specialty, Protective-serv, Tech-support, Transport-moving |

図 2:デモデータ項目と値

これらの属性から資産額を予想するモデルを 2 次利用のデータから構築する場面を想定し、k 匿名化を行ったところ Age 7 -> 2, WorkClass 6 -> 2, Education 12 -> 3, MatritalStatus 4 ->1, Occupation 12 -> 2 と情報量が激減した。

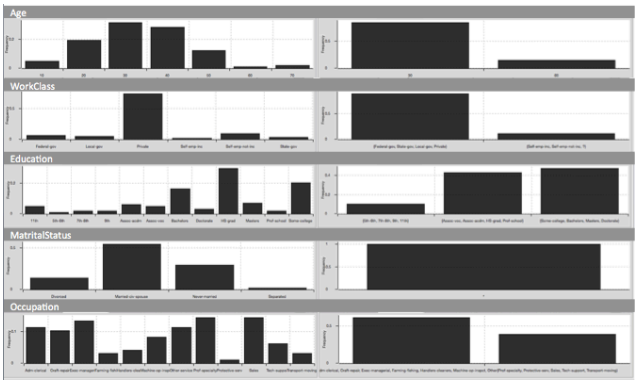


図 3:Incognite アルゴリズムによる k-匿名化

そこから得られた考察としては、準識別子やレコード数が多い場合、従来のフルドメイン一般化では情報損失大きく、データの有用性の担保が困難となる可能性が考えられる。

そこで、2 次利用のための匿名化データの有用性を可能な限り担保しながら匿名化を行う手法が必要と考える。

## 5. 統計的手法による k-匿名化の新手法の提案

2 次データの利用目的として準識別子を説明変数として、機密属性を目的変数とした予測を行う事を想定する。その際の匿名化の効果指標としては予測の精度と置く。

そのため、統計数理研究所の開発したクロス表選択モデルである CATDAP(*Categorical Data Analysis Program Package*)[8]によって、k-匿名化の一般化を行う新手法を提案したい。

CATDAP とはカテゴリーデータである目的変数に対する、説明変数の関連性の度合を AIC(情報量基準)という統計量を用いて解析する多変量解析法であり、以下の 2 種類が存在する。

・AIC を小さい順にならべて変数選択を行う CATDAP-01
・AIC を小さい順にならべてカテゴリーのプーリング(再区分)を行う CATDAP-02

上記の内から、CATDAP-01 を使った k-匿名化を行ったところ有用性が高い(AIC が低い)MatritalStatus の情報量の向上が見られた。

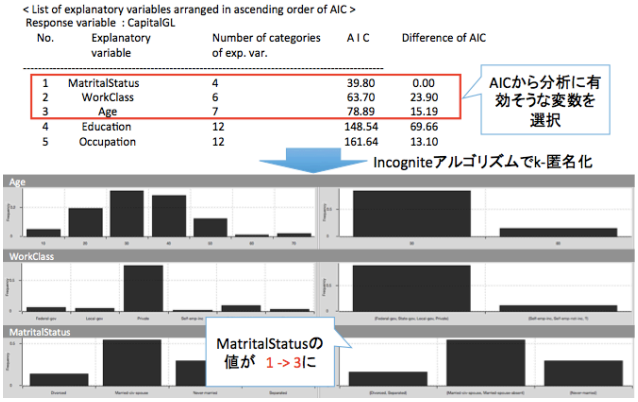


図 4:新手法による k-匿名化

## 6. 考察と今後の課題

目的変数の予測を二次データ利用の目的とする場合、従来手法と比べ、 今回の新手法(ak-匿名化)では一番有用な変数の値数が上がったため、今回のデータにおいては、k-匿名化後の有用性が上がったと考えている。

そのため、実パーソナルデータの匿名加工手法としては、利活用や分析など目的からの k-匿名化のアプローチの可能性を示せた様に考えている。

しかしながら、一般化の方式については今回フルドメイン一般化を使ったため最適解には遠いと考えている。今後はさらにクラスタリングなど、効率的な匿名化アルゴリズムとの組み合わせや、CATDAP-02 によるカテゴリーのプーリング(再区分) による一般化方式の研究が必要と考える。その結果については当日発表を行いたい。

## 参考文献


[1] World Economic Forum, *Personal Data:The Emergence of a New Asset Class*, World Economic Forum, 5 頁, 2011 年.

[2] 佐久間 淳, 小林 重信一, "プライバシ保護データマ イニング ", 人工知能学会誌 V ol24 No2, 2009 年

[3] LATANYA SWEENEY, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002 年

[4] 堀部 政男 他 ,"中間とりまとめ", Suica に関するデータの社外への提供についての有識者会議, 2014 年

[5] Kristen Lefevre , David J. Dewitt , Raghu Ramakrishnan : Incognito: efficient full-domain k-anonymity, (2007)

[6] Kristen Lefevre : Mondrian Multidimensional K-Anonymity, (2006)

[7] Ji-Won Byun1, Ashish Kamra2, Elisa Bertino1, and Ninghui Li1 : Efficient k-Anonymization Using Clustering Techniques , (2007)

[8] 坂元 慶行:**カテゴリカルデータのモデル分析 (応用統計数学シリーズ)**,共立出版,(1985)