～個人情報の安全性と利用性を両立させるために～

k-匿名化可能な属性値の限界区分数予測と個人データ交換のためのk-匿名化処理の提案

小栗 秀暢　　曽根原 登　　国立情報学研究所・総合研究大学院大学

どんな研究？

ビッグデータを安全に利用するためには、情報を匿名化して個人が特定・識別される可能性を低減する技術が不可欠です。

一般的な匿名化技術である「**k-匿名化処理**」は、個人を特定する識別子の数が(k >1)個以上になるよう情報を書き換えたり、クラスタ化する技術です。これにより個人情報の識別可能性を低減させ、目的外利用を抑制します。

本研究では、匿名化処理の結果値を予測するモデルを検討し、そのモデルを活用した個人データ交換システムを提案しています。

適切な匿名化処理パターンを軽量・高速で提供することで、複数回の情報の授受によって発生する情報漏えいリスクを低減させます。

何がわかる？

匿名化処理は、用途に合わせた情報の変更と組み合わせによって多くのパターンが生成されます。 それら全ての情報を厳密にリスク判定・検証するためには、非常に多くの計算リソースが必要となります。

利用者の属性情報は、対象者や退会者数の増減などで常に変化するため、厳密な匿名状態の検定をリアルタイムで行うのは困難でした。

本研究では、実データを用いて、それらの匿名化状態を維持する限界値の推移を研究し、予測モデルを作成しました。 匿名化処理の限界点が高い精度で予測できる場合、厳密な匿名化検定処理を省略できることから、安全性が高い情報を、即時的に利用することが可能となります。

状況設定

・個人情報***P***を匿名化した情報***P'***を公開している***DP***と、その情報を受領してデータ分析対象***Q***と比較・検証したい***DU***が存在する。

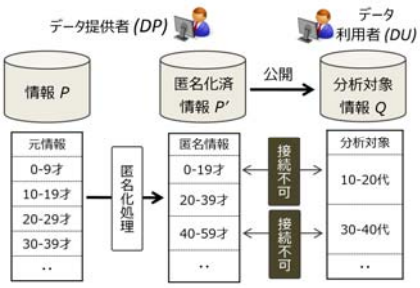
・現在***DP***が公開している情報***P'***と分析対象***Q***は情報の粒度が異なるために比較ができない。

・***DU***が求める粒度でデータの加工を行い、提供して欲しいが、その粒度での匿名化が可能であるかは、処理を試行しないと判明しない。

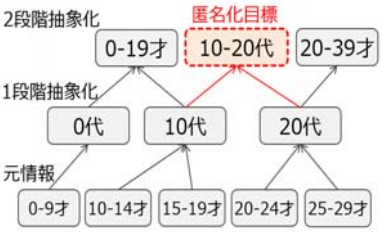
・***DU***の求める粒度で匿名化処理が出来なかった場合、その情報における脆弱なポイント(識別可能性が高い群)が判明する。

・リスクとリソースの両面の問題から、なるべく少ない試行回数で情報の粒度を揃え、両者で合意したい。

■ **データ提供者と利用者の関係性**



■ **情報を抽象化するための判定樹**
匿名化目標が満たせるかを予測する



研究内容

・実データを元に、属性区分数とk-匿名レベルの減少の関係性を調査したところ、累乗近似式(***Y=aX^β***)による予測式が0.99以上の相関を持つことが確認された。

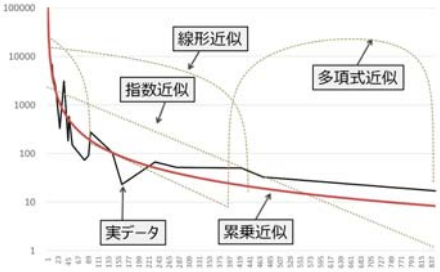
・また、情報の抽象度が高い4～45属性の時点で得られたk-匿名レベルの推移から計算した予測式が、最も精度が高いことが判明した。

・匿名化処理は、属性の組み合わせ数が増加すると計算量が指数的に増加(NP困難)するため、属性種類が少ない内に限界地点を予測することで、計算量を大きく削減できる。

・予測が正確であるならば、不必要な匿名化処理と検証処理を省略することができるため、処理量と情報漏えいリスクの両面において有効である。

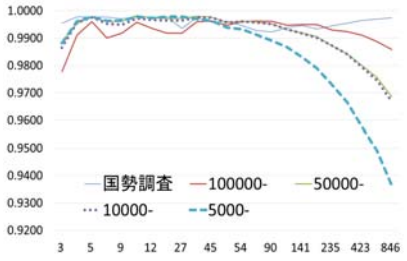
実データと各種近似式の比較

実データは選択肢区分数に従ってｋ＝１に近づく。
主要な近似式の中では累乗近似が最も実データとの相関が高い。

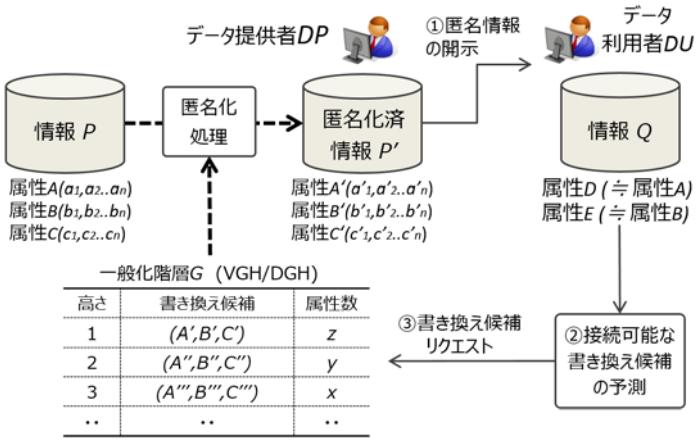


区分数と予測式の相関係数の推移

属性の区分数が少ない(4～45区分)状況で作成された予測式精度が高いことが判明。データが増加すると、特殊値によるノイズで精度が悪化する。



提案する匿名化情報の交換方式



・情報を匿名化して公開しているデータ提供者***DP***と、その保持情報と接続を行うデータ利用者***DU***が互いの持つ情報を公開せずに最適な区分数について合意する仕組みの提案。

・匿名化レベルが予測できない場合、***DU***は***DP***に属性の書き換えリクエストを送り、匿名化情報の検定と評価を繰り返す。

・元情報***A(a1...an)***と公開匿名化情報***A'(a'1... a'n)***，生成される情報***A''(a''n... a''n)***は、情報粒度が異なるため***[an ∈ a''n ∈ a'n]***となる．そのため***a''n***が複数回生成されることで、***an***の情報が類推されるリスクが高まる。

・予測式によって得られた接続区分数が可能である場合のみ、匿名化処理のリクエストを送付する。

	データ提供者DP	データ利用者DU
公開情報	匿名情報(A',B',C') (A',B',C')の生成用一般化階層	
保持情報	元情報(A,B,C)	情報(D,E)
目的	(D,E)と接続可能な粒度で匿名化処理された(A'',B'',C'')の生成	

連絡先：小栗 秀暢　国立情報学研究所 情報学社会相関研究系 曽根原研究室 ／ニフティ株式会社
TEL：03-4212-2731　FAX：03-4212-2120　Email：oguri_hidenobu@nii。ac。jp