
公募シンポジウム

公募シンポジウム8

医療情報の Public Use File提供にむけた検討

2018年11月25日(日) 09:00 ～ 10:30 B会場 (4F 409+410)

[4-B-1-2] 匿名加工コンテスト下の匿名加工技術・有用性指標の開発

○菊池 浩明（明治大学総合数理学部）

2017年の改正個人情報保護法施行により、本人同意がなくとも第三者提供ができる「匿名加工情報」の利用が始まった。2018年には、次世代医療基盤法も施行され、匿名加工情報と匿名加工医療情報を利用した新しいビジネスの創出が期待されている。

しかしながら、匿名加工情報の再識別リスクはそれほど自明ではなく、個人情報保護委員会が公布した「個人情報の保護に関する法律施行規則」第19条の匿名加工情報の作成方法に関する5つの基準は、定性的な記述に留まり、標準的な評価手法は定まっていない。例えば、匿名加工時に想定していた以上の長期間に渡って履歴が流通すると、再識別リスクが高まることが懸念される。そこで、これらの技術的な課題を中手牛、安全な匿名加工アルゴリズムと再識別評価法を競うコンテスト Pws Cupを2015年より開催している。本報告では、2018年10月22日に開催される Pws Cup 2018の結果を報告し、2018年度のテーマである一般化における安全性の基準について考察する。

匿名加工コンテスト下の匿名加工技術・有用性指標の開発

菊池 浩明^{*1}^{*1} 明治大学

Development of Utility and Privacy Metrics of Anonymized Data inspired by Data Anonymizing Competitions

Hiroaki Kikuchi^{*1}^{*1} Meiji University

Abstract

This studies the algorithms of data anonymization and its utility and privacy metrics. It is not trivial to define a common metrics for anonymized data because it is hard to assume a standard adversary model. Hence, the study attempts to explore reliable and acceptable metrics by holding big data competition in which anonymized data are to be re-identified each other.

Keywords: Personal Information Protection Act. Re-identification risk, odds

1. はじめに

2017 年の改正個人情報保護法施行により, 本人同意がなくとも第三者提供ができる「匿名加工情報」の利用が始まった. 2018 年には, 次世代医療基盤法も施行され, 匿名加工情報と匿名加工医療情報を利用した新しいビジネスの創出が期待されている.

しかしながら, 匿名加工情報の再識別リスクはそれほど自明ではなく, 個人情報保護委員会が公布した「個人情報の保護に関する法律施行規則」第 19 条の匿名加工情報の作成方法に関する 5 つの基準は, 定性的な記述に留まり, 標準的な評価手法は定まっていない. 例えば, 匿名加工時に想定していた以上の長期間に渡って履歴が流通すると, 再識別リスクが高まることが懸念される. そこで, これらの技術的な課題を探るため, 安全な匿名加工アルゴリズムと再識別評価法を競うコンテスト Pws Cup を 2015 年より開催している.

本報告では, 2018 年 10 月 22 日に開催される Pws Cup 2018 の結果を報告し, 2018 年度のテーマである一般化における安全性の基準について考察する. そこで, ガイドラインに示された匿名加工医療情報の作成, 提供の考え方を概観するとともに, 今後の具体的な運用を見据えた論点等について整理を行う. 特に匿名加工医療情報の作成については, 改正個人情報保護法の施行規則 19 条を基本としつつ, 追加的な考慮事項が示されており, 国内外の動向等を踏まえて, そのあり方について考察する.

2. 匿名加工コンテスト PWS Cup

2.1 コンテスト概要

表 1 にこれまでに開催された本コンテストの概要を示す. 本コンテストには, 企業の研究所や大学を中心に, 例年 10 以上のチームが参加し, 技術を競っている.

対象としているビッグデータは, 2015 は一世帯一レコードで記述された世帯消費額の疑似データを, 2016 年以降は英国の存在する商取引サイトの購買データを用いている. このデータは, 約 500 名の顧客が年間に 100 件ずつの購買を行った 5 万レコードの履歴を含んでおり, 顧客ごとの差異が大きく, 匿名加工を困難にしている. しかしながら, このような個人情報と紐づけられる履歴データは, 今後匿名加工の対象になるであろう交通機関の乗降履歴やレセプトデータ, 医療履

歴データなど, 多くのデータに当てはまる重要な特徴であり, コンテストとしては避けて通れない.

例年テーマを決めて, 特定の加工方法に絞ったコンテストルールを定めている. 例えば, 2015 年はデータの摂動化などの加工に限定し, 行の削除などは認めていなかった. 2016 年からは, 識別子などの乱数への置き換え, いわゆる, 仮名化を導入した. 改正された個人情報保護法では, 匿名加工情報を, 「個人情報に含まれる特定の個人を識別することができる記述等の全部又は一部を削除すること(当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。)」と定めたことに対応し, この下線が仮名に該当すると考えたためである. ただし, 個人情報保護委員会が定めたガイドライン(匿名加工編)では, 仮 ID について, 「なお, 同じ乱数等の他の記述等を加えた上でハッシュ関数等を用いるなどの手法を用いる場合には, 乱数等の他の記述等を通じて復元することができる規則性を有することとならないように, 提供事業者ごとに組み合わせる記述等を変更し, 定期的に変更するなどの措置を講ずることが望ましい。」と注意を促している. そこで, 2017 年には, 顧客に割当てた仮 ID を途中で変更する加工を認めたルールでコンテストを開催している.

2018 年は, これまでの摂動化などを禁じ, これまではルールに含めていなかった一般化に焦点を当てる.

表 1. 匿名加工コンテスト PWS Cup

	2015	2016	2017	2018
開催	10/21-22 長崎ブリック ホール	10/11-12 秋田キャッスル ホテル	10/23-24 山形国際ホテ ル	10/22 ホテルメトロポ リタン長野
参加者 数	13チーム(20 名)	15チーム(42 名)	14チーム(43 名)	未定
データ セット	疑似ミクロ データ(世帯 消費額)	UCI Dataset "Online Retail" (購買履歴)		
属性数	25	11 (顧客4属性+履歴7属性)		履歴4属性
顧客数	8,333	400	500	500~
履歴数	なし	18,524	44,917	397,625
履歴期 間	1年間	1年間		
加工	摂動化	摂動化, 仮名化, 行削除		一般化

2.2 PWS Cup 2017 から得られた知見

PWS Cup 2017 では、長期的(1年間)な履歴にどのように仮IDを割り当ててくるかが大きな主題であった。有用性の観点では、同じ顧客に一意な仮IDを割り当ててのが望ましい。しかしながら、先のガイドラインでも警告されているように、たとえランダムでも長く同じ仮IDを割り当てていると、識別されやすくなる。そこで、図2の例で示すように、顧客ID(CID)と仮IDの表Fを定め、1から4列目で示される月ごとに別の仮IDを振ることを許している。全ての月について、真の顧客IDと仮IDの対応を正しく予測した時、その顧客は再識別された、と定めた。この例では、1行目と3行目の顧客が再識別され、再識別率は、

$$\text{Reid} = \text{正しい推測顧客数} / \text{全顧客} = 2/5$$

と評価する。この定義については、1年のうちに1か月だけを加工するだけで再識別を免れることになり、異論もあったが、仮名の制御というコンテストの目的が最も効果的に検証できる指標として採用した。

コンテストの結果を図3に示す。ここで、Y軸が再識別率(低い方がよい)、X軸が誤差の大きさ(小さい方がよい)を表し、14チームの分布から戦略の違いが観測できる。左上(図赤丸)は、有用性を高めた(加工幅が小さい)ために、再識別率が高い加工データを提出したチーム群であり、右下(図青丸)はその逆である。このように、有用性と安全性との間にトレードオフが存在することがコンテストの結果から浮かび上がる。優勝したのは、X+Yを最小化する左下のチーム「君の名は〜」である。

仮名の長さ(寿命)と再識別との間には相関はあっただろうか。チームごとに独自の戦略を取っているため、平均的な160件のデータが再識別されたチームを取り上げて解析を行った。図4に、このチームの顧客の再識別(図黒丸)の有無を、各顧客の仮IDの数と仮IDの長さの散布図の上に示す。Y軸は、年間の仮IDの数、X軸は単一の仮IDの平均長(月)を示す。図の左下、すなわち、割当てた仮IDの数が少なく、利用期間の長さも短い顧客は、再識別されやすいことが観測できる。識別された顧客161名の平均仮名長は1.104であり、それは識別されなかった顧客339名の1.223よりも短い。平均仮名長と平均仮名数を説明変数にして、再識別の確率を目標変数としてロジスティック回帰をしたところ、平均仮名数が水準0.01で有意であることが示された。この時のオッズ比(仮名数を1増やすことによる再識別確率の比)は、

$$\text{OR} = e^{-0.46} = 0.63$$

であり、仮ID一つ新たに割り当ててことで、再識別リスクが63%に削減されることを表している。

CID	1	2	3	4
12361	61	61	61	63
12362	62	62	DEL	DEL
12431	31	DEL		31
12628		28		
15100	10	20		40

仮名表 F

CID	1	2	3	4
12361	61	61	61	63
12362	62	77	77	DEL
12431	31	DEL	DEL	31
12628	DEL	28	28	DEL
15100	10	20	DEL	99

推測 ^F

図2. 再識別率の定義

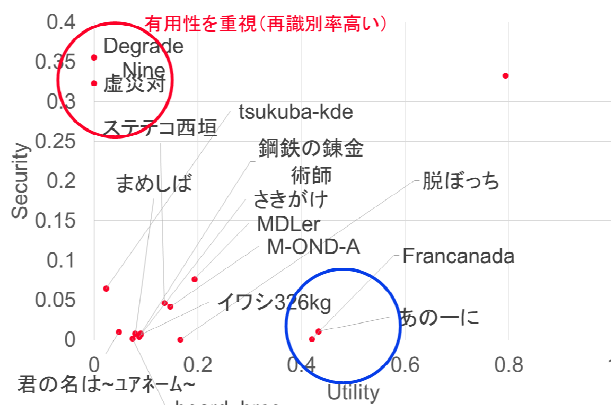


図3. コンテスト結果

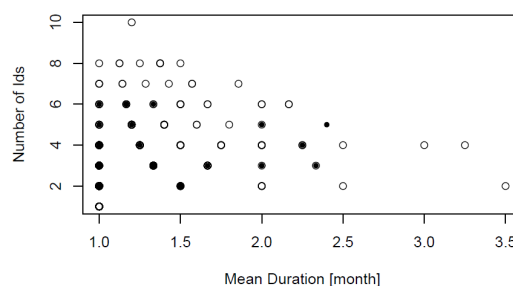


図4. 仮名長と仮名数の分布

2.3 PWS Cup 2018 の狙い

2018年度では、前述した一般化に加えて、安全性の評価方法に大きな工夫を加えている。

2017年度までのコンテストでは、他の参加者が加工されたデータをそれぞれで再識別を行い、その再識別率の最大値や平均値で安全性を見積もっていた。これは定量的で、公平な評価方法である反面、では、どの再識別率までならばその加工データが安全であるかという保証がなかった。どんなに加工を行ったとしても、 n 人の顧客のデータならば、正しく再識別される確率が $1/n$ であり、リスク0にはならない。加えて、購入数が極端に多い顧客や、購入金額が著しく高い顧客がいても、その顧客を削除すると有用性を落としてしまうので、再識別を許して放置した方がコンテストのルール上は有利になってしまう問題もあった。

そこで、2018年度のコンテストでは、顧客集合の部分的な再識別を許すことにした。全顧客を当てて必要はなく、特異な顧客があれば、その顧客だけに絞って再識別をしてもよいルールとしている。再識別する顧客の数に応じて閾値を定め、その閾値を超えたらその加工データは失格とする。従って、先の様な特異な顧客をそのままに放置すると失格となるリスクが増えるので、加工の際に放置されることがなくなることを期待できる。

4. おわりに〜匿名加工医療情報への展望

次世代医療基盤法の下では、一般の事業者に対する個人情報保護委員会規則とほぼ等価な基準に従った加工が示されている。従って、PWS Cupの様なコンテストで得られた有用性や安全性の指標を適用して、匿名加工医療情報のリスクを評価することが考えられる。次世代基盤法では、

認定された事業者により医療情報の照合や匿名加工が行われることになるが、事業者を限定することによりそのリスク評価が偏ってしまわないか懸念される。世界的に、認定された事業者による匿名加工が主流になるとの見方もあり、それらとの整合性が十分に取れるかも気がかりである。本コンテストの知見が、その研究開発に活用されることを願っている。

参考文献

- 1) 個人情報の保護に関する法律についてのガイドライン(匿名加工情報編). 個人情報保護委員会, 2016
- 2) 菊池浩明, 小栗秀暢, 中川裕志, 野島 良, 波多野卓磨, 濱田浩気, 村上隆夫, 門田将徳, 山岡裕司, 山田 明, 渡辺知恵美:PWSCUP2017: 長期間の履歴データの再識別リスクを競う, コンピュータセキュリティシンポジウム CSS 2017, No. 2 (2017)..
- 3) 医療分野の研究開発に資するための匿名加工医療情報に関する法律についてのガイドライン. 内閣府 文部科学省 厚生労働省 経済産業省, 2018

