プライバシーを保護する匿名化技術

Data Anonymization Technique to Protect Privacy

小池　正修　　パキン　オソトクラパヌン　　伊藤　秀将
■KOIKE Masanobu　■Pakin OSOTKRAPHUN　■ITOH Hidemasa

近年，パーソナルデータを利活用して商品やサービスに様々な付加価値を提供するビジネスモデルが増加するなかで，プライバシーを保護するための匿名化技術が注目されている。その中でも*k*−匿名化技術は，個人を特定しようとしても*k*人までしか絞り込めないという特長を持つ有力技術である。その一方で，*k*−匿名化すると，元のデータの多くの値が変更されて情報量が失われるため，データの有用性が低下するという問題があった。

東芝は，従来技術に比べ情報量の損失が約30％少ない*k*−匿名化アルゴリズムを開発した。このアルゴリズムを適用して作成した匿名化データを用いることで，プライバシーを保護しながらより精度の高いデータ利活用を実現できる。

With the ongoing introduction of business models that enhance the added value of products and services and create new value through the use of personal data, anonymization is expected to be a vital technology protecting personal privacy. Among data anonymization techniques, *k*-anonymization is the most promising because a record in a *k*-anonymized dataset is guaranteed not to be linked to an individual with a confidence level of more than 1/*k*. However, the knowledge obtained from a *k*-anonymized dataset is less precise than that in the original dataset due to information loss that occurs during the *k*-anonymization process.

Toshiba has now developed a *k*-anonymization algorithm that reduces the amount of information loss by about 30 % compared with existing techniques. This technique achieves effective utilization of personal data with higher accuracy while protecting personal privacy.

1　まえがき

IoT（Internet of Things）時代では，センサなどのデバイスから収集されたデータを利活用して，様々な付加価値を提供するサービスが重要になる。例えば，携帯電話のGPS（全地球測位システム）機能から収集される位置情報や，オンラインショッピングでの購買履歴などの情報を利用して，その人の状況にあった商品を推薦するサービスなどがある。

このような個人に関わる情報は，パーソナルデータと呼ばれる。パーソナルデータの中には，他人に知られたくないプライバシーに関する情報も含まれるため，利活用する際には，その個人のプライバシーを侵害しない配慮が必要である。

ここでは，利活用したいパーソナルデータは表形式で表されるとし，その各行は1人分のレコードを，各列は属性を表すものとする。例えば**表1**に示す例では，データは7人分のレコードから成り，各レコードは6個の属性を持つ。

パーソナルデータを保有する企業（データ保持者）が，このデータを利活用するために，第三者（データ分析者）に分析を依頼するという状況を考える（**図1**）。従来の個人情報の保護に関する法律（個人情報保護法）では，第三者にデータ分析を依頼することについて，このデータに含まれる全ての人からその旨の同意を得る必要があった。しかし，転居で連絡がつかないなど，全ての人からの同意をとるのが困難な場合が多かった。

表1．表形式のデータ例
Example of data in table format

レコード	属　性					
	名　前	年　齢（歳）	就業区分	学　歴	週間労働時間（h）	年収（$）
1	Alice	50	自営業	大学卒	13	50,000以下
2	Bob	53	民間企業	高校中退	40	50,000以下
3	Charlie	52	自営業	高校卒	45	50,000超
4	Dave	37	民間企業	大学院修士修了	40	50,000以下
5	Ellen	42	民間企業	大学卒	40	50,000超
6	Frank	30	州職員	大学卒	40	50,000超
7	George	32	民間企業	専門学校卒	50	50,000以下

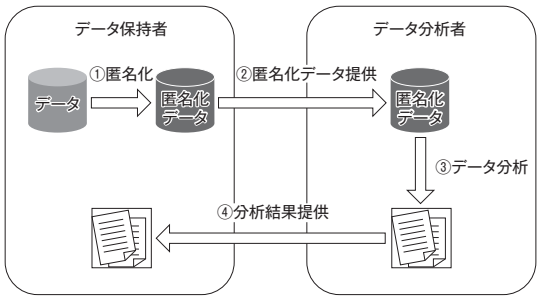


図1．匿名化データの使用例 — データ保持者がデータを匿名化して，データ分析者に提供する。
Example of use of anonymized data

2015年に改正された個人情報保護法では，プライバシーを保護しながら，個人情報の利活用を促進するため，匿名加工情報という概念が導入された。提供される個人情報の項目と提供方法を公表すれば，匿名加工情報を第三者に提供する際に本人の同意は不要になる。匿名加工情報とは，特定の個人を識別できないように加工された情報を指し，パーソナルデータから匿名加工情報を作成する際に用いられる技術が匿名化技術である。

2　匿名化技術

匿名化技術は，データに含まれる名前や住所といった情報を削除したり一般的な値に変更したりするなど，データを加工する技法の総称である。代表的な技法を**表2**にまとめる[1]。

基本的な技法は，氏名や社員番号などそれだけで個人を特定できる属性を削除することである。しかし，データ全体の分析結果を個人に当てはめ，その結果を個人にフィードバックするデータ利活用の方法もある。この場合は，データ保持者が後で個人を特定できるように，仮名化を行う。すなわち，社員番号などのID（識別情報）をランダムかつ一意的な値に置き換える処理である。元のIDとの対応関係は，データ保持者が秘密に保持しておく。

削除や仮名化の対象になる属性は，それ自体をデータ分析のための情報として用いることは少ない。したがって，削除や仮名化をしても，データ分析に与える影響は少ないと言える。その一方，データ分析に使用したい属性でも，特徴的な値である場合は，個人の特定につながることがある。例えば年齢が100歳の人や，世帯員が10人いる人などである。それらの属性を削除すればプライバシーは保たれるが，データ分析結果の精度を損なうおそれがある。そこで，値をある程度一般化した値に置き換えることで，プライバシーとデータの有用性のバランスを取るという匿名化技法が使われる。代表的な技法は，トップコーディングとグルーピングである。

表2．代表的な匿名化技法
Typical data anonymization techniques

技　法	内　容	例
削除	個人を直接特定できる情報を削除	氏名を削除
仮名化	IDなどを別の値（仮名）に置き換える	社員番号を適当な乱数値に置き換える
トップコーディング	極端に大きかったり小さかったりする特殊な属性値を，“○○以上”，又は“○○以下”とまとめる	年齢が85歳以上 世帯員が8人以上
グルーピング	特定の値をグループ分けして，階級区分に変更	22歳⇒20代
リサンプリング	データを全て提供するのではなく，そこから抽出した一部のデータを提供	全体から80％をランダムに抽出して提供
ソート	レコードの配列順を並び替える	－
誤差の導入	レコードの一部の属性に，誤差を加える	身長150 cm⇒150.3 cm

このような技法を用いてデータを加工することで，個人の特定を難しくできる。しかし，表2の技法を用いて匿名化を行ったとしても，いくつかの属性の組合せを調べることで，個人を特定できることがある。例えば，表1で名前の列を削除しても，Ellenが民間企業に勤めていて，かつ大学卒であることを知っていれば，表の中にそのような人は1人しかいないから，5行目のレコードがEllenであることを特定できる。

3　*k*-匿名化

前記の問題を解決するために，*k*-匿名化という概念が提案された[2]。*k*-匿名化は，個人を特定するための絞り込みを行っても，*k*人までしか絞り込めないようにデータを加工することである。*k*を2として表1のデータを*k*-匿名化した例を**表3**に示す。ここでは，年収属性から個人を絞り込むことはできないと仮定し，この属性は加工していない。表3では，年収以外の属性に関し，まったく同じ値の組合せを持つ人が少なくとも2人いるため，*k*すなわち2人未満に絞り込むことができなくなっている。

しかし，そのために表内の多くの値が変更されており，例えば，Aliceの年齢は50歳から50～53歳と幅を持った値へ，就業区分は自営業から有職へと，より広範囲を示す値に変更されている。すなわち，もともとのデータが持っていた情報量が減っており，これは，データの有用性の低下を意味する。

一般に，*k*を大きくすると匿名性は高まるが，情報量の損失が大きくなりデータの有用性が低下するため，匿名性と有用性をいかに両立させるかが*k*-匿名化の大きな課題となっている。しかし，情報量損失のもっとも少ない*k*-匿名化アルゴリズムを理論的に見つけることは困難であるため，実験的に情報量損失のより少ない*k*-匿名化アルゴリズムを見つけるというアプローチがとられている。

k-匿名化アルゴリズムの一つとして，Mondrianが知られている[3]。Mondrianは，処理が高速で，レコード数と属性数に対しスケーラブルであり，実験的に情報量損失が少ないこと

表3．*k*-匿名化されたデータの例（*k*=2）
Example of *k*-anonymized data (*k*=2)

レコード	属　性				
	年齢（歳）	就業区分	学　歴	週間労働時間（h）	年収（$）
1	50～53	有職	中学卒	13～45	50,000以下
2	50～53	有職	中学卒	13～45	50,000以下
3	50～53	有職	中学卒	13～45	50,000超
4	37～42	民間企業	大学卒	40	50,000以下
5	37～42	民間企業	大学卒	40	50,000超
6	30～32	有職	高校卒	40～50	50,000超
7	30～32	有職	高校卒	40～50	50,000以下

一般論文

Mondrian (*dataset*, *k*)

入力データ　*dataset*：　*k*-匿名化の対象のデータセット
　　k：　*k*-匿名化の絞り込み可能人数の下限

出力データ　分割されたデータセット

1. それ以上分割できなければ *dataset* を返す
2. そうでなければ以下を実行
   2-1. *dim* ← 属性を一つ選択する
   2-2. *splitVal* ← 属性 *dim* の値の中央値を求める
   2-3. *lhs* ← *dataset* のうち，属性 *dim* の値が *splitVal* 未満のもの
   2-4. *rhs* ← *dataset* のうち，属性 *dim* の値が *splitVal* 以上のもの
   2-5. Mondrian (*lhs*, *k*) ∪ Mondrian (*rhs*, *k*) を返す

図2．Mondrianアルゴリズム ― 選択した属性でデータを2分割する処理を再帰的に行う。

Mondrian algorithm

が示されているため，多くの論文で使用されている。Mondrianのアルゴリズムを**図2**に示す。

4　東芝の*k*-匿名化アルゴリズム

当社は，Mondrianをベースにして，情報量損失がより少なくなるようにアルゴリズムを改良した[4]。改良のポイントは，データの事前ソートと属性の選び方である。

Mondrianは，選択した属性に関してその中央値でデータを2分割することで，レコードをグループ化していく（図2の処理2-3と2-4）。その際，選択されていない属性も機械的にレコードが2分割されてしまう。そこで，事前に全ての属性に対してレコードをソートして，互いに近い値を持つデータを近くに配置することで，近い値を持つレコードが同じグループに含まれるようにした。

更に，図2の処理2-1で，取りうる値のバリエーションの個数が少ない属性を選ぶことにした。このような属性は，多くのレコードが同じ値を持っていることが期待される。したがって，その属性に対してデータを分割すると，同じ値を持ったレコードが同じグループに分けられる可能性が高い。かりにもともと同じ値を持っているレコードをグループ化できれば，その属性に関しては，情報量の損失はないことになる。

5　評価

当社手法の有効性を評価するために，以下の3種類の実験を行い，Mondrianと比較した。

(1)　損失した情報量の比較
(2)　単純なデータ分析結果の比較
(3)　複雑なデータ分析結果の比較

実験に使用したデータは，UCI（カリフォルニア大学アーバイン校）Machine Learning RepositoryのAdult Data Setである[5]。このデータは，*k*-匿名化の評価の際に事実上の標準として用いられ，レコード数は32,561，属性数は15である。

5.1　損失した情報量の比較

情報量の損失を測る指標として，Loss Metric（*LM*）を使用した[6]。*k*を2，4，8，16，及び32とした場合の*k*-匿名化データについて，Mondrianと当社手法の*LM*を**表4**に示す。

いずれの*k*でも，当社手法のほうが，*LM*が30％程度小さく，情報量の損失をより抑えた*k*-匿名化を実現できる。別の見方をすると，同じ情報量の損失でも，当社手法のほうが*k*をより大きくできるとも言える。例えば，Mondrianで*k*=16とした場合と，当社手法で*k*=32とした場合の*LM*はほぼ同じである。つまり，従来では*LM*=0.18程度の場合は*k*は16が限度であったが，当社手法を用いることで，より匿名性の高い*k*=32を選択することができる。

5.2　単純なデータ分析結果の比較

単純なデータ分析として，年齢属性の度数分布を求めた（**図3**）。当社手法で作成した匿名化データの分布のほうが，元のデータの分布に近い。残差二乗和は，当社手法の場合は387,298なのに対し，Mondrianの場合が1,946,934であり，当社手法の有効性が示された。

5.3　複雑なデータ分析結果の比較

複雑なデータ分析として，年収属性の値（$50,000超か以下かの2値）を他の14属性値から推測する実験を行った。推測

表4．*LM*の比較

Comparison of loss metric (*LM*) values

k	*LM*		比率*
	東芝の手法	Mondrian	
2	0.021	0.027	0.78
4	0.046	0.062	0.74
8	0.079	0.110	0.72
16	0.122	0.176	0.69
32	0.181	0.262	0.69

＊東芝の手法による*LM*をMondrianによる*LM*で割った値

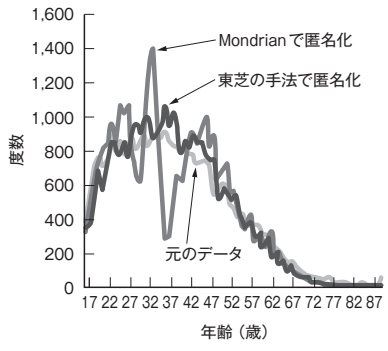


図3．年齢の度数分布 ― 当社手法で匿名化したデータの分布のほうが，元のデータの分布に近い。

Frequency distribution of ages

のために使用したアルゴリズムは，Random Forestである。

推測は，学習フェーズと推測フェーズから成る。学習フェーズ用として，Adult Data Setの32,561件のデータを用いた。推測フェーズでは，Adult Data Setに含まれる推測用の16,281件のデータを用いた。

推測に用いたデータには，実際の年収属性値も記載されており，推測値との答え合わせができる。kが2，4，8，及び16のときに，推測値が実際の値と一致した正解率を示したのが**図4**である。元のデータを使った場合の推測の正解率は0.823であった。匿名化したデータを使った場合の推測の正解率はそれより劣るものの，当社手法はMondrianより多くの正解を導けている。

一方，k=8の場合のROC（Receiver Operation Characteristic）曲線を**図5**に示す。当社手法のほうが，元のデータのROC曲線により近く，元のデータの性質をより残している。また，他のkでも同様の傾向が得られた。

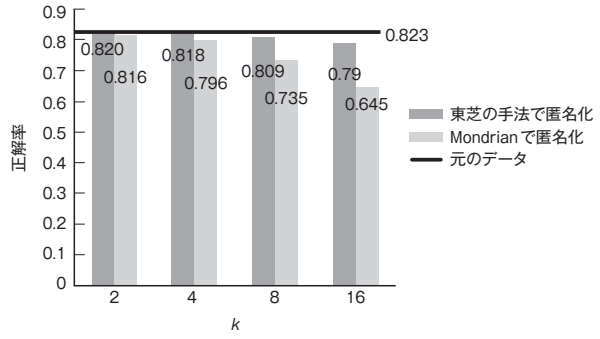
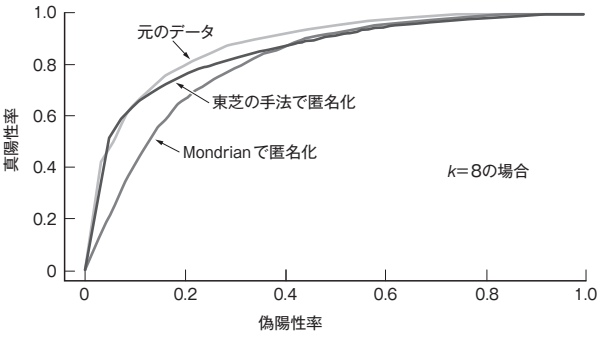


図4．推測の正解率 ― 当社手法で匿名化したデータを用いた場合の正解率のほうが，元のデータでの正解率に近い。

Accuracy rates of estimation



真陽性率：元のデータの年収$50,000超のレコードのうち，$50,000超と推測したレコードの割合
偽陽性率：元のデータの年収$50,000以下のレコードのうち，$50,000超と推測したレコードの割合

図5．ROC曲線 ― 当社手法で匿名化したデータを用いた場合のほうが，元のデータのROC曲線に近い。

Receiver operating characteristic (ROC) curves

6　あとがき

プライバシーを保護しながらデータを利活用するために，匿名化技術が使われている。改正個人情報保護法で匿名加工情報が導入されたこともあり，今後も匿名化技術が使われる場面が増えることが予想される。

匿名化技術の中で，k-匿名化は複数の属性を組み合わせても個人の特定を難しくできる技術である。しかし，k-匿名化により元のデータの情報量が失われることで，データの有用性が低下するという問題があった。これを解決するために当社が開発したk-匿名化手法は，従来技術より情報量損失が少なく，プライバシーを保護しながらより精度の高いデータ利活用を実現できる。

今後は，当社手法のアルゴリズムをいっそう有効なものにするための開発を進め，プライバシーの保護とデータの利活用の両立を図って，社会に貢献していく。

文　献

(1) 総務省．“匿名データの作成・提供に係るガイドライン”．総務省統計局．<http://www.stat.go.jp/index/seido/pdf/35glv4.pdf>，（参照 2015-11-17）．

(2) Sweeney, L. k-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. **10**, 5, 2002, p.557－570.

(3) LeFevre, K. et al. "Mondrian Multidimensional k-Anonymity". 22nd International Conference on Data Engineering (ICDE '06), Atlanta, GA, USA, 2006-04, IEEE. 2006, p.25－35.

(4) Koike, M. et al. "A Mondrian-Based k-Anonymization with Low Information Loss". 2015年暗号と情報セキュリティシンポジウム．北九州，2015-01，電子情報通信学会．2015，3C4-2.

(5) University of California, Irvine. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml/>, (accessed 2015-11-17).

(6) Nergiz, M. E.; Clifton, C. Thoughts on k-Anonymization. Journal of Data & Knowledge Engineering. **63**, 3, 2007, p.622－645.

小池　正修　KOIKE Masanobu, D.Eng.

インダストリアルICTソリューション社 インダストリアルICTセキュリティセンター主査，博士（工学）。情報セキュリティ技術の研究・開発に従事。情報処理学会会員。

Industrial ICT Security Center

パキン　オソトクラパヌン　Pakin OSOTKRAPHUN

インダストリアルICTソリューション社 インダストリアルICTセキュリティセンター。情報セキュリティ技術の研究・開発に従事。電子情報通信学会会員。

Industrial ICT Security Center

伊藤　秀将　ITOH Hidemasa

インダストリアルICTソリューション社 IoTテクノロジーセンター データ利活用技術開発部。ビッグデータ処理技術及び機械学習技術の研究・開発に従事。情報処理学会会員。

IoT Technology Center