情報処理学会論文誌 Vol.56 No.12 2244–2252 (Dec. 2015)

データベース分割再構成法による k-匿名化が誘発する濡れ衣の軽減

角野 為耶[1,a)] 荒井 ひろみ[2] 中川 裕志[2]

受付日 2015年3月9日, 採録日 2015年9月2日

概要：個人の情報を保護したデータ開示法の 1 つに k-匿名化がある．k-匿名化されたデータを人間が閲覧した際に，データに含まれた人間に対して不利益を生ずるような推測がなされる場合がある．本研究ではこの現象を k-匿名化が誘発する濡れ衣と呼び，濡れ衣を発生させうる属性を持つ機微なレコードに着目し，濡れ衣の発生を軽減させる k-匿名化法を提案する．実データに対して濡れ衣を発生させうる機微属性を付与したデータセットを用いて実験を行い，提案手法を用いると濡れ衣を軽減させた k-匿名化を実現できることを確認した．

キーワード：k-匿名化，データベース，濡れ衣，プライバシ，パーソナルデータ

Preventing A False Light Caused by k-anonymity with Dividing Databases

Nasuka Sumino[1,a)] Hiromi Arai[2] Hiroshi Nakagawa[2]

Received: March 9, 2015, Accepted: September 2, 2015

Abstract: In the field of privacy preserving data mining, k-anonymity is a representative model for protecting privacy. However, when people see k-anonymized data, a person who provides his/her data is misleadingly suspected as a bad guy due to the information which actually has nothing to do with him/her. We define such problem as a false light caused by k-anonymization, and define a record which has an attribute causing a false light as a sensitive record. We propose k-anonimization algorithms which pay attention to sensitive records in order to prevent a false light. In the experiments, we confirmed that proposed method can decrease a probability of occurrence of a false light.

Keywords: k-anonymity, database, false light, privacy, personal data

1. はじめに

昨今，ビッグデータの利活用に注目が集まっており，とりわけ個人の行動や属性について記録したパーソナルデータはマーケティング等における利用が期待されている．一方で，個人に関する機微な情報が含まれているためにパーソナルデータの活用は個人のプライバシの侵害につながりうる．このような事態を解決するために個人のプライバシを保護しつつデータを活用する技術の研究が進められている．その代表的な技術の 1 つに 2002 年に提案された k-匿名化 [4] があり，k-匿名化の発展については Fung らの論文 [2] にまとめられている．しかし，k-匿名化はデータに含まれている人間が不利益な推測をなされる濡れ衣という問題を誘発する恐れがある [11]．本研究では，既存の k-匿名化法が濡れ衣が発生する可能性を考慮していないことに着目し，濡れ衣が発生する可能性を抑える k-匿名化アルゴリズムおよび k-匿名化されたデータの濡れ衣の誘発を評価する指標を提案する．

[1] 東京大学大学院学際情報学府
Graduate School of Interdisciplinary Information Studies, The University of Tokyo, Bunkyo, Tokyo 113–0033, Japan
[2] 東京大学情報基盤センター
Information Technology Center, The University of Tokyo, Bunkyo, Tokyo 113–8658, Japan
[a)] nasuka.sumino@gmail.com

2. k-匿名化と濡れ衣

2.1 k-匿名化

パーソナルデータの属性は個人 ID，擬似 ID，機微情報およびその他の 4 種類に分類できる．個人 ID とは氏名等の単体で個人の特定につながるような属性であり，擬似 ID とは年齢や職業等の単体では個人の特定につながらないが複数組み合わせることによって個人の特定につながりうるような属性である．機微情報とは病名等の他人に暴露された場合に問題となる属性である．パーソナルデータの例を**表 1** に示す．Sweeney は擬似 ID と外部情報の突き合わせによって個人が特定される危険性を指摘し [5]，個人の特定を防ぐための手法として k-匿名化を提案した [4]．k-匿名化とはデータが持つ個人 ID を消去したうえで擬似 ID の情報の一部を消去（抑圧）あるいは精度を落とし（一般化），同じ擬似 ID を持つ人間がデータベース内に k 人以上存在するようにデータを変換する手法である．このようなデータ変換によって，外部情報と突き合わせたとしても個人を特定できる確率を $\frac{1}{k}$ 以下に抑えることになる．表 1 のパーソナルデータを 4-匿名化したデータの例を**表 2** に示す．表 2 の例においては，一般化によって年齢が 30 歳から 39 歳という値の範囲に置き換えられており，住所は文京区という範囲の広い地域に置き換えられている．本研究では表 2 における 4 つのレコードのようにまったく同じ擬似 ID を持つレコード群を匿名化グループ，もしくは単純にグループと呼ぶ．

ここで病歴という属性について考えてみる．風邪のような軽度な病歴は一般的には漏洩しても本人に不利益はないであろう．しかし肝炎等の長期にわたって健康状態に影響を与える病歴は就職活動中の学生等にとっては漏洩すると不利益になる恐れがある．機微情報の中でも特にこのような不利益をもたらす可能性があるデータを機微データと呼ぶ．

表 1 パーソナルデータのデータベース例
Table 1 A data base example of personal data.

名前	年齢	性別	住所	病名
一郎	39	男	文京区本郷	肝炎
次郎	35	男	文京区湯島	肝炎
三子	33	女	文京区弥生	風邪
四郎	30	男	文京区本駒込	風邪

表 2 4-匿名化の例
Table 2 An example of 4-anonymous data.

名前	年齢	性別	住所	病名
A	30-39	*	文京区	肝炎
B	30-39	*	文京区	肝炎
C	30-39	*	文京区	風邪
D	30-39	*	文京区	風邪

k-匿名化は 1 つの匿名化グループに存在する機微情報の種類に制約をかけていない．そのため，1 つの匿名化グループ内に肝炎の人間しか存在しない場合，そのグループに属する個人が肝炎であることが暴露されてしまう．この問題を解決するために，k-匿名化を行う際に 1 つの匿名化グループ内に存在する機微情報の種類を ℓ 種類以上にする制約を設ける ℓ-多様性という概念が提案されている [1]．表 2 においては 2-多様性が実現されている．

パーソナルデータに記載される属性には数値的な属性およびカテゴリカルな属性の 2 種類の属性が存在する．前者は年齢や身長等の数値で示される属性であり，後者は住所や職業等の数値で示せない属性である．実用的な k-匿名化アルゴリズムを提案するには，これらの 2 種類の属性に対して適用できるアルゴリズムを設計する必要がある．

2.2 k-匿名化が誘発する濡れ衣

本節では k-匿名化が誘発する濡れ衣という問題の概要を説明する．表 2 において 4-匿名化を行ったことによって誰が肝炎であるのか区別がつかなくなり，肝炎を患っていない人間が肝炎であることを疑われる恐れがある．k-匿名化によって機微データを持つと疑われる現象を濡れ衣と呼ぶことにする．濡れ衣は機微データを持つレコードが匿名化グループに含まれた場合に発生する．肝炎のような濡れ衣を発生させうる機微データを含むレコードを本研究では機微なレコードと呼ぶ．

2.3 濡れ衣が発生するメカニズム

本節では濡れ衣が発生する原理について，新卒採用における意思決定の例を用いて説明する．中川らは匿名化グループに機微なレコードがどの程度存在すれば濡れ衣が発生するかを主観確率を用いて説明している [11]．企業で採用担当をしている人物 A とその企業の採用に応募している人物 B がいると仮定する．B を含む k-匿名化された医療データが公開されており，A がそれを閲覧したとする．B が含まれる匿名化グループ中で肝炎を患った人間の割合が高い場合，B は実際には肝炎でなくても肝炎であることを強く疑われる恐れがあり，企業は B ではなく別の人物を採用する等の対策をとる可能性がある．このような場合は別の人物の採用にあたって面接等のコストをかける必要があり，こうした対策にかかるコストを対策コストと定義する．企業側の観点では，B をそのまま採用した場合は肝炎によって仕事に支障が生じる等の損失が発生する恐れがあり，この損失を被害額の期待値と定義する．被害額の期待値は匿名化グループ中の肝炎の人間の割合に比例するが対策コストは一定であり，被害額の期待値が対策コストを超えた場合に A は対策をとることが合理的だと考えられる．よって，A が B を肝炎であると疑う主観確率は，被害額の期待値が対策コストを超えない場合は 0 となり，対策

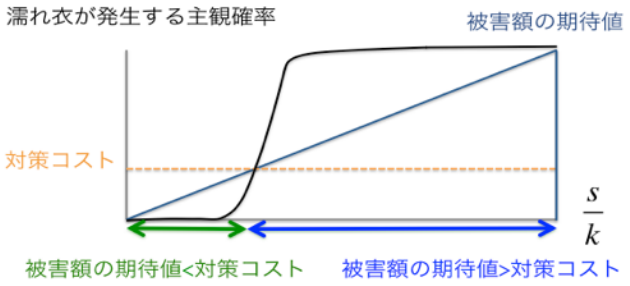


図 1 濡れ衣が発生する主観確率
Fig. 1 Subjective probability of false light occurring.

コストを超えた場合に急上昇して 1 となる，単位ステップ関数のような曲線を描くと考えられる．しかし，現実世界では肝炎の可能性が低くても疑いを持ち対策を実行する人もいれば可能性が高くても楽観視して対策を行わない人もいる．したがって，肝炎の人間がどの程度の割合で匿名化グループに存在すれば主観確率が 1 となるかは，A がどのような人物かによって異なる．このような不確実性を考慮し，本研究では主観確率の曲線を**図 1** のような曲線で近似することにした．ここで k は匿名化グループに存在するレコード数であり，s は匿名化グループ中に存在する機微なレコード数，このケースにおいては肝炎を患っている人数である．

図 1 のような濡れ衣が発生する主観確率の曲線の具体的な関数として，式 (1) のようなシグモイド関数を採用する．α は主観確率の不確実性を示すパラメータであり，β は匿名化グループにおいて機微なレコードがどの程度の割合で存在すれば濡れ衣が発生するかを示すパラメータである．

$$p\left(\frac{s}{k}\right) = \frac{1}{1+e^{-\alpha\left(\frac{s}{k}-\beta\right)}} \tag{1}$$

濡れ衣は ℓ-多様性を満たす匿名化グループにおいても発生する．表 2 のような 2-多様な匿名化を行った場合であっても，機微なレコードが匿名化グループ中に存在する割合が 50%と高い値であるため，濡れ衣が発生する主観確率は非常に高い値となる．ℓ-多様性は匿名化グループ中に存在する機微情報の種類に制約を付けることはできるが，機微なレコードの割合までは制御できないため，図 1 のような主観確率を考慮して濡れ衣を防ぐことはできない．したがって，機微なレコードのグループ内での比率に着目して濡れ衣を防ぐという本研究の提案には整合しない．

2.4 濡れ衣を軽減する k-匿名化の方針

濡れ衣の発生を防いで k-匿名化を行うには 2 つの方針が考えられる．第 1 の方針は k-匿名化の k を大きくする方針である．匿名化グループにおいて濡れ衣が発生する主観確率は $\frac{s}{k}$ に依存し，この割合が一定値を超えると急激に上昇する．したがって，k を増加させて分母を大きくすることによって図 1 の曲線に示した濡れ衣が発生する主観確率が上昇することを防ぐことができる．たとえば，k=10 の匿名化グループの中に 5 つ機微なレコードが存在すれば濡れ衣が発生する主観確率は高い値になるが，k=50 であれば匿名化グループ中の機微なレコードの割合は 10 分の 1 であり，濡れ衣が発生する主観確率を抑えることができる．ただし k を大きくするとデータの精度が低下するという問題が発生する．

表 3 濡れ衣を考慮した k-匿名化の例
Table 3 An example of false light aware k-anonymous data.

名前	年齢	性別	住所	病名
B	26-35	*	文京区	肝炎
C	26-35	*	文京区	風邪
D	26-35	*	文京区	風邪
E	26-35	*	文京区	風邪
F	26-35	*	文京区	風邪

第 2 の方針は機微なレコードを別々の匿名化グループに割り当てて 1 つの匿名化グループに存在する機微なレコードの割合を小さくする方針である．機微なレコードを 5 つとして，この 5 つのレコードを k=10 の 5 つの匿名化グループに割り当てる．このように機微なレコードを分割して割り当てることで，各匿名化グループに含まれる機微なレコードの割合を 10 分の 1 に抑えることができ，かつ k を小さく保てるのでデータの損失を抑えることができる．

本研究では第 2 の方針を採用し，k-匿名化において濡れ衣が発生する主観確率を最小限に抑えるために k-匿名化法を提案する．表 1 のデータを k-匿名化した場合の例を**表 3** に示す．表 2 の匿名化グループ中にはもともと機微なレコードが複数存在していた．本研究では，機微なレコードが複数存在している匿名化グループに対して，機微なレコードを別のグループに割り当て，機微でないレコードを別の匿名化グループから取り入れることで濡れ衣が発生する主観確率を最小限に抑えた k-匿名化を実現する．

3. k-匿名化の評価指標

本章では k-匿名化アルゴリズムの評価指標として以下に示す濡れ衣が発生する主観確率，k-匿名化による情報損失を示す指標である NCP，機械学習におけるクラス分類の精度を評価する AUC について説明する．

3.1 濡れ衣が発生する主観確率の評価指標

本研究では k-匿名化において濡れ衣が発生する主観確率を評価する指標を新たに提案する．K 個の匿名化グループから構成される匿名化グループの集合を $C = \{C_1, C_2, \ldots, C_K\}$ とし，その添字集合を $I = \{1, 2, \ldots, K\}$ とする．この匿名化グループで構成される匿名化されたデータベース T' において濡れ衣が発生する主観確率を式 (2) で定義する．

$$S_{\max}(T') = \max_{i \in I}\left\{\frac{1}{1+e^{-\alpha\left(\frac{s_i}{|C_i|}-\beta\right)}}\right\} \tag{2}$$

この指標は，k-匿名化によって生成された匿名化グルー

プの中で最も濡れ衣を誘発しやすいグループがどの程度濡れ衣を誘発するかを測るものである．

3.2 情報損失

k-匿名化を適用したことによる属性およびデータベースの情報損失を測る指標として，Xu らが提案している NCP (Normalized Certainty Penalty) [7] を用いる．以下のように数値的な属性の NCP とカテゴリカルな属性の NCP を定義する．

3.2.1 数値的な属性の NCP

本項では数値的な属性の NCP について説明する．k-匿名化されたデータにおいては，レコードが持つ数値的な属性の値はそのレコードがその属性値においてとりうる値の範囲で表現される．m 個の数値的な属性 $(A_1, A_2, \ldots, A_m)$ を持った N 個のレコードからなるデータベース T を考える．データベース T 中の 1 つのレコードを $t_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,m})$ として定義し，このレコードを k-匿名化によって精度を落としたレコードを $t'_i = ([y_{i,1}, z_{i,1}], [y_{i,2}, z_{i,2}], \ldots, [y_{i,m}, z_{i,m}])$ とする．ただし，$x_{i,j}$ はデータベース T 中のレコード t_i における属性 A_j の属性値であり，$y_{i,j} \leq x_{i,j} \leq z_{i,j}$ かつ $1 \leq j \leq m$ である．このとき，t'_i の 1 つの属性 A_j に関する数値的な属性の NCP は以下のように定義される．

$$NCP_{A_j}(t'_i) = \frac{z_j - y_j}{|A_j|} \tag{3}$$

ここで，$|A_j|$ は属性 A_j がデータベース中でとりうる最大値から最小値を引いた値である．次に 1 つのレコードのすべての属性の NCP を定義する．式 (3) を用いて，レコード t_i の NCP は式 (4) のように定義される．ただし，w_j は属性 A_j に対してその重要性に応じて重み付けをするパラメータであり，$0 \leq w_j \leq 1$ かつ $1 \leq j \leq m$ である．

$$NCP(t'_i) = \Sigma_{j=1}^{m} w_j \cdot NCP_{A_j}(t'_i) \tag{4}$$

3.2.2 カテゴリカルな属性の NCP

本項ではカテゴリカルな属性の NCP について説明する．カテゴリカルな属性は階層構造を持つ．カテゴリカルな属性では階層構造を利用し，属性の値を階層構造のより上位のノードの値で置換することによって一般化を実現する．カテゴリカルな属性の NCP は階層構造において属性値がどの程度一般化されているかを考慮して計算される．**図 2** のような，階層化されたカテゴリカルな属性 A_j を考える．階層構造におけるあるノード u の $size$ を，そのノードが持つ葉ノードの数として定義する．図 2 の例では $size(v_1)$=2，$size(v) = 7$ となる．匿名化によって，カテゴリカルな属性 A_j の値を一般化した値 u_j を持つレコード t'_i を考える．このとき，レコード t'_i の属性 A_j の NCP は以下のように定義される．

$$NCP_{A_j}(t'_i) = \frac{size(u_j)}{|A_j|} \tag{5}$$

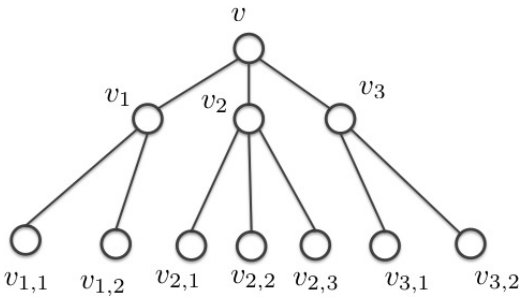


図 2 カテゴリカルな属性の階層構造の例

Fig. 2 An example of hierarchical structure of categorical attributes.

ただし，$|A_j|$ は属性 A_j が持つ葉ノードの数である．属性 A_j の値が v_1 に一般化されているレコード t'_i があると仮定すると，このレコードの属性 A_j の $NCP_{A_j}(t'_i)$ の値は $\frac{2}{7}$ となる．次に，1 つのレコードのすべての属性に関する NCP の計算方法を示す．m 個のカテゴリカルな属性 $(A_1, A_2, \ldots, A_m)$ を持った N 個のレコードからなるデータベース T を考えると，このデータベースのあるレコード t'_i のカテゴリカルな属性の NCP は式 (6) のように定義される．w_j は数値的な属性の場合と同様に，属性 w_j の重要性に応じて重み付けをするパラメータである．データベースに存在する各属性 A_j に対して，式 (5) を用いて計算した $NCP_{A_j}(t'_i)$ とパラメータ w_j の重み付き和が 1 つのレコードのカテゴリカルな属性の NCP となる．

$$NCP(t'_i) = \Sigma_{j=1}^{m} w_j \cdot NCP_{A_j}(t'_i) \tag{6}$$

3.2.3 匿名化データ全体の NCP

カテゴリカルな属性と数値的な属性を合わせて m 個持つ k-匿名化されたデータベース T' を考える．このデータベース全体の NCP は式 (7) で定義される．ただし，$NCP_{A_j}(t')$ は属性 A_j の形式に合わせて計算する．Xu らの文献 [7] ではデータベースに存在する全レコードの NCP の総和を匿名化されたデータベースの NCP としていたが，本研究では全レコードの NCP の総和をレコード数で割って正規化した値をデータベース全体の NCP と定義する．

$$NCP(T') = \Sigma_{t' \in T'} \Sigma_{j=1}^{m} \frac{w_j \cdot NCP_{A_j}(t')}{|T'|} \tag{7}$$

3.3 k-匿名化データを用いたクラス分類の予測性能

k-匿名化がデータの利活用にどの程度影響を与えるかを評価するために本研究では k-匿名化データを用いたクラス分類を行う．k-匿名化データを用いたクラス分類については Inan らの先行研究 [8] がある．先行研究を参考に本研究では k-匿名化された訓練データから分類器を学習し，k-匿名化が施されていないテストデータに対してクラス分類の性能を評価する．先行研究ではクラス分類の性能評価に正解率を用いているが，クラス分類においては実験に用いるデータセット中の各クラスに属するデータ数に偏りがある

Algorithm 1 匿名化グループの再構成

```
1: C = {C_i | i = 1, 2, ..., K}：初期匿名化で匿名化された匿名化グループの集合
2: TC ← φ：再構成対象のグループの集合
3: while 機微なレコードの割合が 1/k を超える匿名化グループが存在する do
4:    C_* ← C 中で最も機微なレコードの割合が高い匿名化グループ
5:    C ← C \ C_*
6:    TC ← near_groups(C_*, C)
7:    h ← |TC| − 1
      //匿名化グループを再構成した際に出力するグループ数．再構成対象となったグループ数から一つ減らした値を再構成後のグループ数とする
8:    C から near_groups で選ばれたグループを取り除く
9:    RC ← re_anonymize(TC, h)
10: end while
11: for i = 1 to |RC| do
12:    C ← C ∪ RC_i
13: end for
```

Algorithm 2 $near\_groups(C_*, C)$

```
1: TC ← φ：再構成対象のグループの集合
2: s(TC)：再構成対象のグループ中に含まれる機微なレコード数
3: m：再構成対象となったグループ数
4: TC ← TC ∪ C_*
5: m ← 1
6: while s(TC) > m − 1 do
7:    d ← arg min_i (NCP(C_*, C_i ∈ C))
8:    if C_d 中に機微なレコードが含まれている then
9:       C_d を near_groups の探索対象から除外する
10:      continue
11:   else
12:      TC ← TC ∪ C_d
13:      m ← m + 1
14:   end if
15: end while
16: return TC
```

場合があり，このような場合は正解率を用いても適切に分類性能を評価することができない．そこで本研究では，偏りのあるデータの分類問題において学習器を評価する手法として用いられる AUC（Area Under the Curve）[9] で予測性能を評価する．AUC は ROC 曲線の下の部分の面積のことであり，ROC 曲線とは分類器のパラメータを変化させた際の偽陽性率 vs 陽性率における評価値の列である．AUC を用いることでパラメータを変化させた際の分類器の総合的な性能を評価できる．

4. 提案手法

本研究では，濡れ衣の発生を抑えたうえで情報損失を最小化する k-匿名化アルゴリズムを提案する．一般的な k-匿名化アルゴリズムでは情報損失の最小化を目指す場合が多いが，本研究では情報損失に加えて濡れ衣を考慮する．k-匿名化における濡れ衣の発生を最小限に抑えるために，1 つの匿名化グループに存在する機微なレコードの割合を $\frac{1}{k}$ 以下に抑えるという制約を設け，そのうえで情報損失の最小化を目指す k-匿名化アルゴリズムを提案する．

4.1 濡れ衣を軽減させ匿名化グループの再構成

本研究で提案する k-匿名化アルゴリズムは初期匿名化および匿名化グループの再構成の 2 つのプロセスで構成される．初期匿名化では既存のアルゴリズムを用いてデータベースを k-匿名化する．匿名化グループの再構成では初期匿名化によって生成された匿名化グループの集合を入力とし，1 つの匿名化グループに存在する機微なレコードの割合が $\frac{1}{k}$ 以下である匿名化グループの集合を出力する．再構成アルゴリズムの流れを Algorithm 1 に示す．

再構成アルゴリズムは初めに全匿名化グループから機微なレコードが存在する割合が最も高いグループ C_* を再構成対象の匿名化グループとする．次に $near\_groups$ を用いて C_* の周囲のグループを再構成対象の匿名化グループとして選択する．選択されたグループ中のレコード群から，1 つのグループに機微なレコードが存在する割合が $\frac{1}{k}$ 以下になるよう匿名化グループの再構成を行う．機微なレコードに関する制約に違反した匿名化グループが存在しなくなるまで以上の処理を行うことで，濡れ衣が発生する主観確率を最小限に抑えた k-匿名化を実現する．

4.2 再構成対象グループの選択

Algorithm 2 の $near\_groups$ は，再構成に用いる匿名化グループを選択するアルゴリズムである．このアルゴリズムは初めに再構成対象とするグループの集合 TC を空集合として初期化した後に C_* を追加し，再構成対象のグループ数 m を 1 とする．次に，再構成対象のグループを選択するために C_* から最も距離が近い匿名化グループ C_d を探索する．匿名化グループ間の距離は，2 つのグループに属するすべてのレコードの擬似 ID がすべて同一になるような最小限の一般化を行った際の NCP を匿名化グループ間の距離とする．C_d 中に機微なレコードが存在しない場合は TC に C_d を追加し，機微なレコードが存在する場合は C_d を追加せず次に距離が近いグループを探索する．以上の処理を Algorithm 2 の 6:の条件を満たすまで繰り返す．このような場合分けを行うことで，C_d の周囲に機微なレコードの制約に違反した匿名化グループが密集している場合に再構成対象のグループを過剰に追加し続けることを防ぐ．6:の繰返しの条件は $re\_anonymize$ で出力する匿名化グループの数と再構成対象のグループの集合中に存在する機微なレコードの数を同数にするためにこのように設定した．繰返しが終了した後，再構成対象の匿名化グループの集合 TC を戻り値として返す．

4.3 濡れ衣を軽減させる再構成

Algorithm 3 の $re\_anonymize$ は，再構成対象の匿名化

Algorithm 3 $re\_anonymize(TC, h)$

```
RC = {RC_i | i = 1,2,...,h}:再構成される匿名化グループの集合
r = {r_i | i = 1,2,...,|r|}:再構成対象となった匿名化グループに存在するレコードの集合
各匿名化グループ RC_i ∈ RC を空集合に初期化
for i = 1 to h do
    匿名化グループ RC_i に対して機微なレコードを一つ割り当てる
    5:で割り当てた機微なレコードを r から取り除く
end for
各匿名化グループ RC_i が持つレコード数が k 個になるまで 9:〜13:の操作を繰り返す
for i = 1 to |RC| do
    d ← arg min_j (NCP(RC_i ∪ r_j ∈ r))
    RC_i ← RC_i ∪ r_d
    r ← r \ r_d
end for
for j = 1 to |r| do
    e ← arg min_i (NCP(r_j ∪ RC_i ∈ RC))
    RC_e ← RC_i ∪ r_j
    r ← r \ r_j
end for
return RC
```

グループの集合 TC を入力とし，機微なレコードに関する制約に違反しない匿名化グループの集合 RC を出力する．初めに再構成後の各匿名化グループ RC_i を空集合に初期化し，Algorithm 3 の 4:から 7:の処理で各匿名化グループ RC_i に 1 つずつ機微なレコードを割り当てていく．$near\_groups$ において，TC 中に存在する機微なレコード数と $re\_anonymize$ で出力する匿名化グループの数を一致させるように繰返し文の条件を設定しているため，以上の処理が終了した時点で再構成対象のグループ TC 中に存在していたすべての機微なレコードを別々の匿名化グループに割り当てることが可能となり，機微なレコードに関する制約に違反しない匿名化グループを生成できる．次に，各匿名化グループ RC_i に対してレコード数が k 個になるまで RC_i と距離が最も近いレコードを割り当てていく．匿名化グループとレコード間の距離は，グループに属するすべてのレコードと対象のレコードの擬似 ID がすべて同一になるような最小限の一般化を行った際の NCP を匿名化グループ間の距離とする．その後，残ったレコードを最も距離が近い匿名化グループに割り当てる．このような 2 段階の割当てプロセスを経ることによって，情報損失が大きな匿名化グループの発生を防ぐ．$re\_anonymize$ では入力したグループ数よりも 1 つ少ない数の匿名化グループを出力するが，これは情報損失の増加を防ぐためである．匿名化グループ内の機微なレコードを 1 つ以下にするという目的だけなら匿名化グループ数を 1 つ以上を減らすこともありうる．しかしそれでは情報損失が大きくなるので，情報損失を不必要に増加させないために入力した匿名化グループ数より 1 つ少ない匿名化グループを出力することにした．

なお，提案手法においては前述のように機微なレコードに関する制約を設けて匿名化グループを再構成するため，匿名化グループの数に対して機微なレコードの数があまりにも多い場合は提案手法を適用することができない．すべてのレコードを匿名化グループに割り当てた後，再構成された匿名化グループの集合 RC を戻り値として返す．

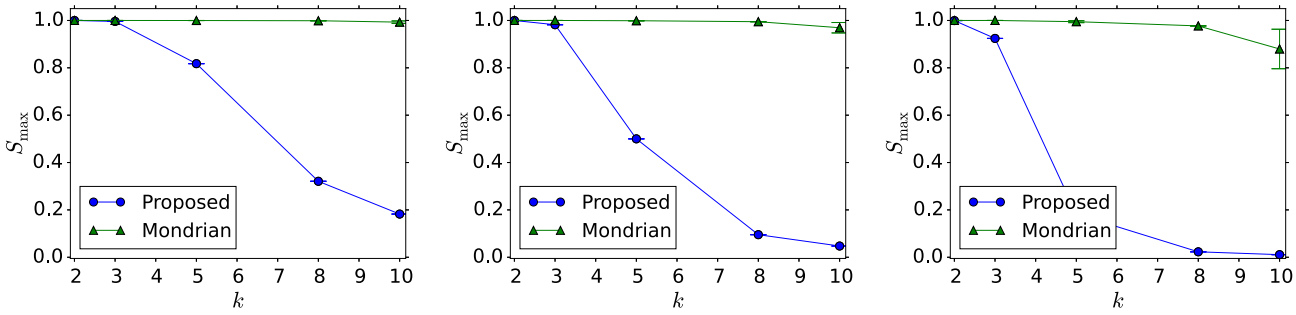
5. 評価実験

5.1 実験データ

評価実験には，UCI の Adult Dataset および pima Dataset を用いた．Adult Dataset は，k-匿名化の研究において広く実験に用いられているデータセットであり，年齢，教育年数の 2 つの数値的な属性と職種，業種，性別，人種，出身国，結婚歴の 6 つのカテゴリカルな属性を持つ．先行研究 [6] に倣って欠損値を持つデータを除去し，30162 個のレコードを k-匿名化の評価実験に用いた．pima Dataset は年齢，BMI，妊娠回数，血糖値，拡張期血圧，皮下脂肪，血清中インシュリン，遺伝的要素の 8 つの数値的な属性を持つデータセットであり，768 個のレコードを k-匿名化の評価実験に用いた．濡れ衣が発生する主観確率を評価するためにそれぞれのデータセットに対して機微なレコードであるか否かを示す属性を追加した．機微な属性の追加にあたって，一様乱数を用いて 4%の確率で機微なレコードであるという情報を付与し，残りのレコードに対しては機微なレコードではないという情報を付与した．機微なレコードの割合は問題依存であるが，今回は試験的に 4%に設定した．機微なレコードの割合が高くなるほど機微なレコード数が増えるため，機微なレコードの割合が高くなると提案手法は機能しない．実験にあたっては機微な属性の付与の乱数の初期値が異なる 5 つのデータセットで実験を行い，その平均を記録した．乱数の初期値によって実験結果が異なるため，結果のグラフには信頼度 95%区間を示すエラーバーをプロットした．NCP の計算において各属性を重みづけするパラメータ w_j はすべて 1 に設定した．また，濡れ衣に関わるパラメータ α は 30 とした．これらのデータセットに対して既存手法および提案手法で k-匿名化を施し，それぞれの手法の有用性を濡れ衣が発生する主観確率，NCP，AUC で評価した．

5.2 初期匿名化法

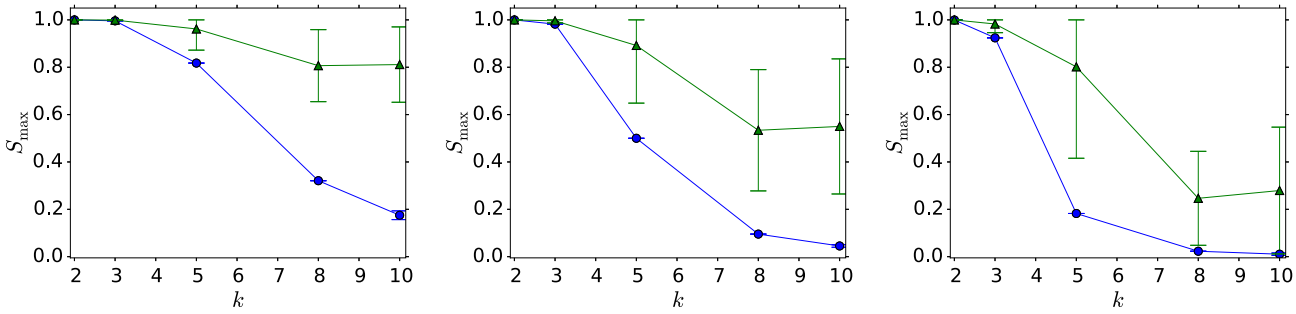
実験を行うにあたって，提案手法における初期匿名化には Mondrian [3] を用いた．Mondrian は kd 木 [10] を用いて，トップダウン的に匿名化グループを生成する．初めに，すべてのレコードのすべての属性を完全に一般化した状態にし，1 つのパーティションとする．このパーティションに対して属性を 1 つ選択し，その属性において各要素がパーティション中に何回出現しているかカウントする．数値的な属性の場合は出現回数に対して中央値を $splitVal$ として



(a) Adult: β=0.15 の濡れ衣誘発指標　(b) Adult: β=0.20 の濡れ衣誘発指標　(c) Adult: β=0.25 の濡れ衣誘発指標

図 3　Adult Dataset の濡れ衣誘発指標の k，β に対する変化

Fig. 3　Subjective probability of false light occurring for various values of k and β in Adult Dataset.



(a) pima: β=0.15 の濡れ衣誘発指標　(b) pima: β=0.20 の濡れ衣誘発指標　(c) pima: β=0.25 の濡れ衣誘発指標

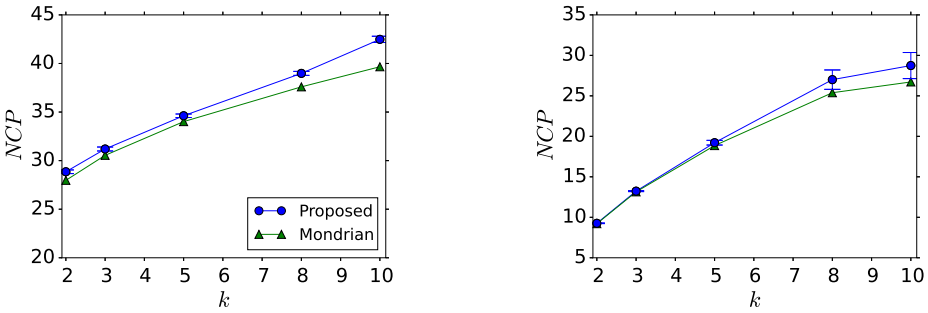
図 4　pima Dataset の濡れ衣誘発指標の k，β に対する変化

Fig. 4　Subjective probability of false light occurring for various values of k and β in pima Dataset.

選択し，その属性において $splitVal$ 以下の値を持つレコード群と $splitVal$ より大きい値を持つレコード群の 2 つのパーティションに分割する．カテゴリカルな属性の場合はパーティションにおける子ノードの出現回数を加算していき，出現回数の合計がレコード数の半分を超えた時点でそれまでに加算された子ノードのレコード群と加算されていない子ノードのレコード群に分割する．パーティション内の要素数が k 個未満となるまで再帰的にパーティションを分割し，すべてのパーティションが分割できなくなった段階で各パーティションを匿名化グループとして出力する．

5.3　濡れ衣が発生する主観確率の評価

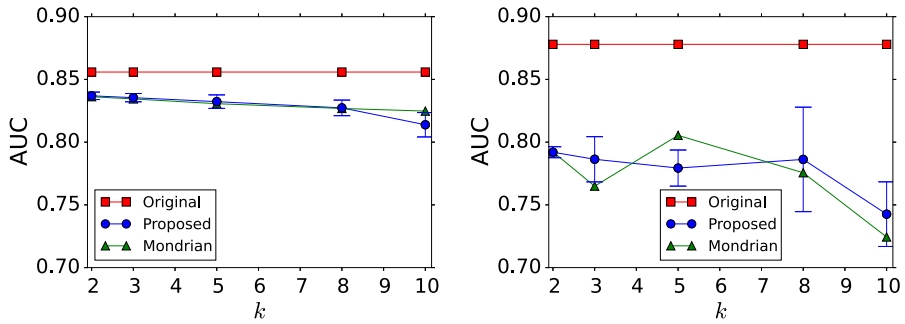
式 (2) を用いて k=2, 3, 5, 8, 10，α=30，β=0.15, 0.2, 0.25 と変化させた場合の濡れ衣が発生する主観確率を評価した．また，紙面の都合上掲載は省くが，α=10, 50 の場合においても同様の実験を行った．α を変更した場合においても，k の変化に対しする主観確率の推移は同様の傾向を示した．Adult Detaset の実験結果を図 **3** (a)，(b)，(c) に，pima Dataset の実験結果を図 **4** (a)，(b)，(c) に示す．実験の結果，どちらのデータセットも k=5, 8, 10 の場合においては既存手法と比較して提案手法が良い結果を残した．提案手法はすべてのグループに対して機微なレコードが存在する割合を $\frac{1}{k}$ 以下にする制約をかけているため，機微なレコードの割合が最も高い匿名化グループにおいても濡れ衣が発生する主観確率を低く抑えられる．一方で，Mondrian では制約がないために機微なレコードが多数含まれる匿名化グループが存在する．k の値を増やした場合でも，図 1 の曲線の右側で値をとっている匿名化グループが存在するため，このような結果になったと考えられる．k=2, 3 の場合は提案手法においても $S_{\max}$ が大きな値となっている．これは，匿名化グループに存在するレコード数が少ない場合，1 つでも機微なレコードを含むと濡れ衣が発生する主観確率が最大となってしまうことに起因する．k が小さい場合は匿名化グループに含まれる機微なレコードが 1 つであっても濡れ衣が発生する主観確率は図 1 において曲線が上昇しきった 1 に近い値をとる．提案手法を用いた場合でも機微なレコードはいずれかの匿名化グループに 1 つは含まれるため，k が小さい場合は濡れ衣が発生する主観確率が非常に高くなる．pima Dataset においては β=0.25, k=8, 10 の場合は Mondrian でも濡れ衣が発生する主観確率が低い値をとっている．pima Dataset は Adult Dataset と比較してデータセットが小規模であり，匿名化グループ数が少なくなる．そのため Mondrian を用いても機微なレコードを多数含む匿名化グループが出現しにくくなり，濡れ衣が発生する主観確率の最大値である $S_{\max}$ が低い値に抑えられていると考えられる．



(a) Adult Dataset の情報損失指標の k に対する変化　(b) pima Dataset の情報損失指標の k に対する変化

図 5　情報損失指標の k に対する変化

Fig. 5　Information loss for various values of k.



(a) Adult Dataset の AUC　(b) pima Dataset の AUC

図 6　k に対する AUC

Fig. 6　AUC for various values of k.

5.4　情報損失の評価

k=2, 3, 5, 8, 10 と変化させた際の提案手法および Mondrian を適用した場合の NCP を評価した．結果を図 **5** (a), (b) に示す．Mondrian を適用した場合は乱数の初期値による結果の変動がないため，エラーバーをプロットしていない．実験より，提案手法は Mondrian にわずかに劣る結果となった．提案手法では，1 度 Mondrian で作成した匿名化グループに対して制約を付けて部分的に再構成しているため，再構成した匿名化グループの NCP が増加し，このような結果になったと考えられる．k が増加するほど NCP が増加しているが，これは k の値が増加することによってより多くのレコードを一般化して匿名化グループを形成する必要があるからだと考えられる．

5.5　匿名化データを用いたクラス分類

匿名化された各データセットの訓練データで分類器を学習し，匿名化されていないテストデータを二値分類する実験を行った．Adult Dataset では 30162 件の訓練データで学習を行い，15060 件のテストデータにおいて年収が 5 万ドル以上か否かを予測する二値分類を行った．pima Dataset では 500 件の訓練データで学習を行い，268 件のテストデータにおいて糖尿病か否かを予測する二値分類を行った．学習アルゴリズムには RBF カーネルを用いた SVM を使用し，SVM の正則化パラメータ C は 1 とした．

5.6　AUC

クラス分類における AUC に関する実験結果を図 **6** (a), (b) に示す．図中では匿名化を施していないオリジナルデータで学習した結果を Original と表記し，提案手法および Mondrian で k-匿名化を施したデータで学習した結果をそれぞれ Proposed，Mondrian と表記している．Mondrian およびオリジナルデータは乱数の影響を受けず，分散がないためエラーバーはプロットしていない．実験の結果，それぞれのデータセットにおいてオリジナルデータで学習した分類器が最も良い結果を残した．Adult Dataset においては，提案手法，Mondrian ともに同程度の精度となり，また k の値が増加するほどに分類の精度が落ちる結果となった．k-匿名化によってレコードが持つ属性が一般化されたために，このような結果になったと考えられる．pima Dataset においては提案手法のエラーバーの範囲が Adult Dataset と比較して広くなっている．これは Adult Dataset と比較して pima Dataset のレコード数が少ないために，機微属性を付与する際の乱数が k-匿名化で出力するデータに与える影響が大きいからだと考えられる．

6.　おわりに

本研究では濡れ衣を軽減する k-匿名化手法を提案した．提案手法では，1 度 k-匿名化したレコードに対して濡れ衣が発生する主観確率を最小限に抑えるように匿名化グループを再構成した．実験の結果，提案手法は既存手法と比較

して情報損失の点では若干劣るものの，濡れ衣が発生する主観確率を大きく軽減できることが確認できた．k-匿名化したデータを用いたクラス分類の実験では，サイズが大きなデータセットを用いた実験では既存手法と同程度の性能を記録し，サイズが小さなデータセットを用いた実験でもほとんどの k の値で同程度の性能を記録した．濡れ衣の発生によってパーソナルデータを提供した人間が様々な不利益を被る恐れがあり，既存手法による k-匿名化では濡れ衣は避けて通れない重大な問題である．データベースを k-匿名化し，かつ濡れ衣による不利益を解消するには提案手法が最も優れた手法だといえる．

参考文献

[1] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubaramaniam, M.: L-diversity: Privacy beyond k-anonymity, *Journal ACM Trans. Knowledge Discovery from Data* (*TKDD*), Vol.1, No.1 (2007).

[2] Fung, B.C.M., Wang, K., Chen, R. and Yu, P.: Privacy-preserving data publishing: A survey of recent developments, *Journal ACM Computing Surveys* (*CSUR*), Vol.42, No.4 (2010).

[3] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Mondrian multidimensional K-anonymity, *Proc. International Conference Data Engineering*, (*ICDE*), p.25 (2006).

[4] Sweeney, L.: k-Anonymity: A Model for Protecting Privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol.10, No.5, pp.557–570 (2002).

[5] Sweeney, L.: Uniqueness of simple demographics in the US population, Technical Report, Carnegie Mellon University (2000).

[6] Iyengar, V.S.: Transforming data to satisfy privacy constraints, *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.279–288 (2002).

[7] Xu, J., Wang, W., Pei, J., et al.: Utility-based anonymization using local recoding, *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–790 (2006).

[8] Inan, A., Murat, K. and Elisa, B.: Using anonymized data for classification, *Proc. IEEE 25th International Conference on International Conference Data Engineering* (*ICDE*), pp.429–440 (2009).

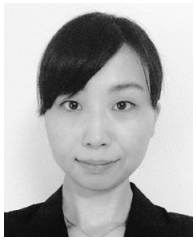
[9] Jin, H. and Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms, *IEEE Trans. Knowledge and Data Engineering*, Vol.17, No.3, pp.299–310 (2005).

[10] Friedman, J., Bentley, J. and Finkel, R.: An algorithm for finding best matches in logarithmic time, *ACM Trans. Mathematical Software*, Vol.3, No.3, pp.209–226 (1977).

[11] 中川裕志，角野為耶：滞在場所の k-匿名化と濡れ衣，情報処理学会第 62 電子化知的財産・社会基盤研究発表会（EIP 研究会），Vol.2013-EIP-62, No.12 (2013).

角野 為耶

1989 年生．2013 年東京大学卒業．2013 年同大学大学院入学．濡れ衣に配慮した k-匿名化の研究に従事．



荒井 ひろみ

1981 年生．2010 年東京工業大学大学院博士（理学）取得．筑波大学研究員，理化学研究所基礎科学特別研究員を経て，2014 年より東京大学情報基盤センター助教．機械学習，プライバシ保護技術の研究に従事．

中川 裕志 （正会員）

1953 年生．1975 年東京大学卒業．1980 年同大学大学院博士課程修了．横浜国立大学を経て，1999 年より東京大学情報基盤センター教授．自然言語処理，統計的機械学習，プライバシ保護データマイニングの研究に従事．