様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月　 ９ 日現在

機関番号：１２６０１
研究種目：挑戦的萌芽研究
研究期間：2014～2015
課題番号：２６５４００４１

研究課題名（和文）匿名化が誘発する濡れ衣現象のモデル化と最適化による解消法

研究課題名（英文）Preventing A False Light Caused by k-anonymity with Mathematical Modling and Optimization

研究代表者

中川　裕志（NAKAGAWA, Hiroshi）

東京大学・情報基盤センター・教授

研究者番号：２０１３４８９３

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：個人の情報を保護したデータ開示法の一つに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するk-匿名化法を提案する。滞在位置情報、およびカテゴリー情報も含む一般的な場合に対して、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したk-匿名化を実現出来ることを確認した。

研究成果の概要（英文）：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-anonimisation algorithms which pay attention to sensitive records in order to prevent a false light. We deal with two cases: location information preservation and more general information including categorical information. In the experiments, we confirmed that proposed method can decrease a probability of occurrence of a false light.

研究分野：情報工学

キーワード：プライバシー　k-匿名化　濡れ衣　個人情報　パーソナルデータ　位置情報　属性情報

１．研究開始当初の背景

ビッグデータにおいてデータベースに記載された個人が特定される問題は 2000 年以降大いに指摘された。例えば、病院の患者データベースにおいて個人名を消去ないしは無意味な番号への仮名化により匿名化しても、年齢、住所、性別などの属性情報（これらを疑似ＩＤと呼ぶ。）が知られてしまうと、それらを利用して、個人名が記載された他のデータベース（米国の場合は公開が義務付けられている選挙人名簿）と付き合わせることによって、個人が識別されうる。結果として個人の病歴など深刻な情報（機微な情報ともいう。）が特定され、流出する可能性がある。その対策として Sweeny[1]によって提案された k-匿名化は、事象 S の当事者が一意に識別されないように、事象 S と関係する当事者を含むデータベースにおいて疑似ＩＤの精度を落とすなどの改変を加え、同じ疑似ＩＤの組み合わせを持つ人が k 人以上いるような曖昧化を図る方法である。

ビッグデータのうちその価値の高さが注目され、利用が促されているのはパーソナルデータとりわけ個人の行動履歴である。特に地理的行動履歴は、公共交通の設計や商品販売戦略などのビジネスに直結し役立つと言われている。個人の行動履歴情報については、その利用に関して漠然とした危惧をいだく人が多数であろう。例えば、2013 年 7 月に JR 東日本が Suica の行動履歴データを匿名化した上で日立製作所を経由して販売するという報道がされたとたんに、反対意見や多くの懸念が噴出し、販売計画は中止されている。したがって、匿名化をした場合の問題点を明確にし、対策を示すことがビッグデータ活用の要となる。これまでは匿名性の問題に焦点が当てられ、地理的行動履歴に対して個人の滞在した精密な位置情報を k 人以上その中に滞在する大きさの領域に精度を落とすことによって k-匿名化を行った。ところが、この k 人以上滞在する領域内に病院、整形クリニック、消費者金融、銃砲店、裁判所、風俗営業店などの施設がある場合、こういった施設に行かなかった人も施設に出入りしたことが疑われること、すなわち濡れ衣の疑いがかかる可能性がある。仮に就職活動中あるいは婚活中の人物であれば、このような濡れ衣によって大きな不利益を被りかねない。この問題点は従来のプライバシー保護データマイニングの分野では全く指摘されてこなかった。

行動履歴以外にも個人データは数多く、購買履歴、職業、収入など多岐にわたる。これらについても同様にプライバシー保護と同時に濡れ衣を防ぐ方策が必要になる。

２．研究の目的

本研究では、個人の地理的滞在履歴などにおける k-匿名化に焦点を当て、この匿名化がかえって濡れ衣の疑いを誘発してしまう仕組みを数理モデル化する。次に濡れ衣の疑いの発生によって生ずる損害とデータマイニングの精度低下による損失の総和を最小化する最適化問題として濡れ衣の問題の解消法を定式化する。最適化問題は解析的に解けないので、シミュレーションによって提案手法における最適化問題の解の実験的評価を行う。

３．研究の方法

(1)人の地理的行動の履歴、とりわけ滞在場所に関する情報をビッグデータとして用いる場合は、k-匿名化されたデータベースへの検索によって利用される。この匿名化が誘発する濡れ衣の被疑の主観確率のモデル化を行う。次に、(a)濡れ衣の被疑による損害と(b) k-匿名化で滞在場所を曖昧化しデータマイニング精度が低下したことによる損失の両者を総合した目的関数を定義する。この目的関数を経済的利害得失として定義する。
(2)個人の滞在場所情報に関して、上記の目的関数を最適化する。具体的には、k-匿名化における領域の分割方法を種々に変動させて最適化問題として定式化する。
(3)さらにカテゴリー情報を含む一般的情報の場合にも、上記の目的関数最適化を行うデータベース分割方法に関しても最適化問題として定式化する。
(4)上記の(2)(3)項によって定式化された最適化問題を解くアルゴリズムを開発、実装し、具体的データを使って評価実験を行う。

４．研究成果

(1)主観確率による濡れ衣発生メカニズムの解明

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。企業で採用担当をしている人物A 及び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 は実際には肝炎ではなくとも肝炎であることを強く疑われる恐れがあり、企業側はB ではなく別の人物を採用する等の対策を取る可能性がある。このような場合は別の人物の採用にあたって面接や調査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越えた場合にA は対策を取ることが合理的だと考えられる。よって、B がA を疑う主観確率は、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、被害額の期待値が対策コストを超え

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。しかしながら、現実世界では肝炎の罹患率が低くとも疑いを持ち対策を実行する人もいれば、罹患率が高くとも楽観視して対策を行わない人もいる。従って、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、B がどのような人物なのかによって異なる。本研究ではこのような不確実性を考慮し、主観確率の曲線を図1のような曲線で近似する。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重要な病名などの機微なレコード数、このケースにおいては肝炎を患っている人数である。

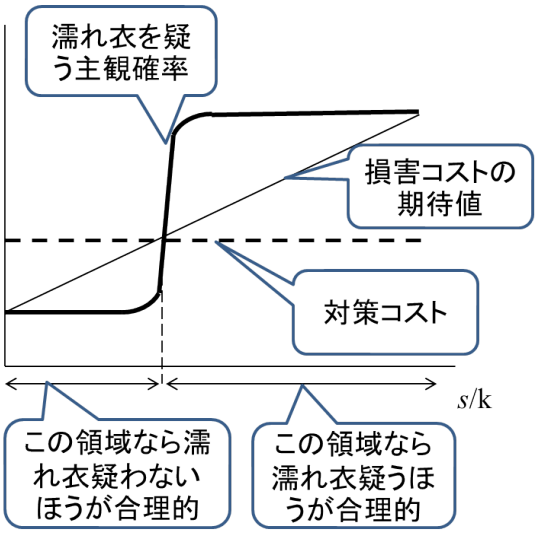


図１　対策コスト、損害期待値と主観確率の関係

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近似関数としてシグモイド関数を採用した。

$$p(s/k)=1/(1+\exp(\alpha(s/k-\beta))) \qquad (1)$$

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。

(2)濡れ衣を軽減するk-匿名化

濡れ衣の発生を防ぎ、かつk-匿名化を行うには２つの方法が考えられる。第一の方法はk-匿名化のkを大きくすることである。匿名化グループにおいて濡れ衣が発生する主観確率は、上述のように匿名化グループに存在する機微なレコード数s を匿名化グループ全体のレコード数k で割った値に依存し、この割合が一定の値を越えると急激に上昇する。従って、kの値を増加させて分母を大きくすることによって、図1 の曲線における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出来る。しかし、kを大きくするとデータの精度が落ち、利用価値が低下する。もう一つの方法は、機微なレコード数sをできるだけ小さくすることである。理想的には、s=1 が望ましい。kの値はプライバシー保護の要請により決められるものである。よって、sを1に近づけるためには、k-匿名性を保ったまま、匿名化グループの構造を変化させることである。

次に考えなければいけないのは、匿名化グループの構造をどのような方向に変化させるかである。これは、匿名化グループの構造変化を最適化問題と捉えれば、目的関数の設定法ということになる。k の値が予め決まっているなら、プライバシー保護の性能は、ある個人が類似の k 人に紛れ込むことは保証されている。すると、匿名化グループの変更においては変更後のデータの情報損失をできるだけ低くすることが目的になる。このために以下のような関数を用いる。

m 個の数値的な属性$(A_1, A_2,\ldots,A_m)$ を持ったN 個のレコードから成るデータベースT を考える。データベースT 中の一つのレコードを$t_i = (x_{i1},x_{i2},\ldots, x_{im})$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$t'_i=([y_{i1}, z_{i1}],\ldots, [y_{im}, z_{im}])$ とする。ただし、$y_{ij} \le x_{ij} \le z_{ij}$　ただし、$1 \le j \le m$ である。このとき、t'_i　の一つの属性A_j に関する数値的な属性の情報損失 は以下のように定義する。

$$\text{情報損失}(t_i)=\sum_{j=1}^{m}\left(z_{ij}-y_{ij}\right)\Big/|A_i| \qquad (2)$$

ただし、$|A_i|$は属性 A_i の最大値－最小値である。

一方、カテゴリ属性の場合は、レコードをtに含まれるカテゴリー属性Aのある属性値vを1個に属性値に一般化した値とき、各々に含まれる属性値の種類数をsize(A)、size(v)とするとき、

情報損失(t)=　(size(v)-1) /size(A)　　(3)

とする。一例を図２に示す。

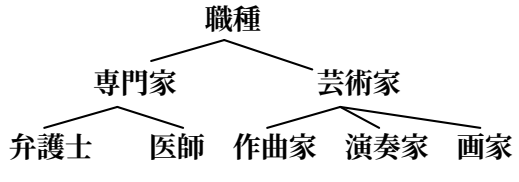


図2　カテゴリー属性

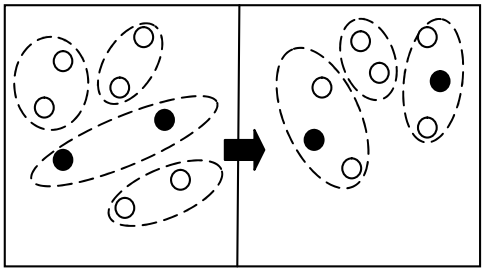
size(職種)は、弁護士、医師、作曲家、演奏家、画家の7個である。同様にsize(専門家)=2、size(芸術家)=3である。ここで、作曲家、演奏家、画家を一般化して芸術家とすると、式(3)により

情報損失(職種)
=(size(芸術家)-1)/size(職種)=2/7
となる。
以下では、個々人の滞在位置を対象にする場合と、数値情報や職位、病名などのカテゴリー情報を含むより一般的な場合の 2 種類について、上記の方針で濡れ衣発生を抑制するアルゴリズムを提案する。

(3)滞在場所情報における濡れ衣発生の軽減方法
まず、機微なデータ点となる個人を定義する。これは、消費者金融ショップ、特殊な病院など機微な場所に滞在した個人として定義する。
k-匿名化はトップダウン型で行い、具体的には K-平均法を用いる。個人は 2 次元の領域に滞在するので、x,y 座標の組で表現できる。したがって、k-匿名化によって k 人を含むグープは 2 次元平面上の領域となる。まず、全体で N 人いるデータの場合は、K-平均法の重心の数の初期値は N/k 個となる。各個人の滞在場所と重心の位置の差はユークリッド距離を使う。

1:初期配属段階
K-平均法では、データの各点が最も近い重心に付随するようにグループ化される。ただし、あるデータ点を最近傍の重心のグループに附属させようとしたとき、(a)既にそのグループに k 人入っている場合、(b)既に機微なデータ点が入っている場合、その重心のグループには所属させず、次に近い重心のグループに所属させる。

2:グループ再構成段階
再構成プロセスでは、まずグループ内の情報損失を計算し、式(2)で定義した情報損失が大きいグループから順に再構成の対象としていく。対象となったグループの周囲L 個のグループを取り込み、再構成対象のグループと取り込んだグループのデータ点から新たにL 個のグループを形成する。L はパラメータである。
初めに再構成対象になったグループとその周囲L 個のグループを再構成の対象としているため、インプットとなるグループ数は$L+1$ 個である。しかしこれらのグループから$L+1$ 個のグループを出力すると再び情報損失の大きなグループが生成されるため、グループ数を1 つ減らしてL 個のグループを再構成する。つまり、なお、再構成に関わるグループ中にL 個以上の機微なデータ点が含まれている場合は一つのグループに複数の機微なデータ点が割り当てられて濡れ衣が発生するため、取り込むグループの数を増加させることで濡れ衣の発生を防ぐ。例えば、4 グループから 3 グループへの再構成は図 3 のようになる。



●が機微なデータ点である。
図 3　グループ再構成

評価実験
以上のアルゴリズムを 100x100 の大きさの領域に 405 個のデータ点をランダム配置し、機微な場所を 2 カ所とし、機微な場所から距離 6 以内に入ったデータを機微なデータ点とした。式(1)のα=30、β=0.25 として、既存手法(参考文献[2]、[3])と比較した。既存手法は k-匿名化はしているが、濡れ衣発生確率の低減はしていないので、濡れ衣発生確率は各々k=5 において 0.6 程度、k=10 で 0.9 程度であった。一方、提案手法では濡れ衣発生確率は k=5 で 0.18、 k=10 で 0.01 であった。問題となる情報損失については、k=5 で提案手法は既存手法より 50%上昇したが、k=10 では 10～20%の上昇に抑えられた。なお、計算時間は K-平均法の場合、大半の時間が 1: 初期配属段階で費やされているため、2:グループ再構成段階の処理を追加しても有意な増加は観測されなかった。よって、濡れ衣発生確率を抑える処理による計算時間の目立った増加は発生していないといえる。

(4)一般的情報における濡れ衣発生の軽減方法
個人データが数値情報やカテゴリー情報が混在する一般的な場合は、まず既存の手法で k-匿名化を行う。その上で機微なデータ点、すなわち問題のある病名の診断を受けている個人、あるいは問題のある職業の人など、が k 人からなるグループに複数存在する場合はグループ再構成を行う。すなわち、基本的なアルゴリズムの枠組みは、上記(3)滞在場所情報の場合と同じである。
グループ再構成の方法について以下に述べる。

再構成アルゴリズムの大枠
1: 全グループの中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を再構成対象のグループとして選択する。
2:Cの周囲のグループを再構成対象の匿名化グループ集合として選択する。
3:選択されたグループ集合を再構成して機微なデータが一つのグループに複数割り当てられていないグループの集合を出力する。
4:以上の処理を、C中に機微なレコードが複数割り当てられているグループが存在しなくなるまで続けることによって、濡れ衣が発

生する主観確率を抑えたk-匿名化を実現する。

再構成すべきグループ集合の選択法

1:グループ間の距離は、グループの中心同士の距離で測る。距離とは、１データに含まれる属性値の距離である。カテゴリー属性の場合は、属性値が木構造ないしグラフ構造をなすので、グループ中心の属性値を繋ぐ最短経路の長さとする。

2:C に近接するグループを集めるとき、以下の 2 点を考慮する。

2-1: 近接するグループのうち、C と合併したとき、元のグループの場合に比べて情報損失の増え方が最少のグループから順に集めていく。なお、合併するとは、疑似 ID が同じになるように疑似 ID の精度を粗くすることである。

2-2: 集合内の機微なデータ数が（グループ集合数－１）より大きくならない最少な数グループを集める。

3: 上記 2：で集めたグループ集合に対して、機微データが 1 個以上入らないように k-匿名化のグループ分けを行う。ただし、グループ数を減らす場合は元のグループ数より 1 以上減らないようにする。これは、情報損失を抑えるためである。

評価実験

評価実験には、k-匿名化 の性能評価で用いられるUCI のAdult Data Set[4] を用いた。Adult Data Set は年齢等の数値的な属性及び性別等のカテゴリカルな属性を持ち、30162 個のレコードで構成される。評価実験では数値的な属性を2 種類、カテゴリー属性を6 種類選択してデータを抽出し、抽出したデータ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5% の確率で機微なデータ点であるという情報を付与し、残りは機微なデータ点ではないという情報を付与した。

k-匿名化はMondrian法[5]を用いた。α=30、β =0.25でk = 3, 5, 8, 10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。

主観確率の平均値は、Mondrian法、提案手法ともすると、kが大きくなるにつれて減少し、k= 3で0.1程度であったものがk= 5でMondrianが0.4、提案手法が0.3、k=10でMondrian、提案手法とも0.1以下になった。一方、主観確率が最大の場合は、Mondrianではk= 3, 5, 8, 10と変化させても0.9程度であったのに対し、提案手法では、k=3で0,9 がk=5で0.2、k=10で0.01以下になった。

情報損失は、Mondrian法と提案手法はほぼ同じであった。

以上の結果から、提案手法は、情報損失は増加させずに、濡れ衣発生確率をk=5以上で0.1以下に減少させることに成功した。よって、k-匿名化が誘発する濡れ衣発生という副作用を抑え、かつデータ有用性も保つことができる有力な方法であることが実験的に示された。

参考文献

[1] Sweeney L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 10.05:557-570. 2002

[2] Gruteser M and Dirk. Anonymous usage of location-based services through spatial and temporal cloaking.*Proceedings of the 1st international conference on Mobile systems, applications and services* . 2003

[3] Mokbel M. F, Chi-Yin C and Walid G. A. The new Casper: query processing for location services without compromising privacy.*Proceedings of the 32nd international conference on Very large data bases* . 2006.

[4] UCI repository of machine learning databases, https://archive.ics.uci.edu/ml/datasets/Adult.

[5] LeFevre K, DeWitt D. j, and Ramakrishnan R. : Mondrian multidimensional K-anonymity. *In Data Engineering, Proceedings of the 22nd International Conferenceon*. pp.25-25. 2006.

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① 中川 裕志、パーソナルデータの利活用における技術および各国法制度の動向編集にあたって．情報処理学会誌 査読無、Vol.55,No.12、2014、pp.1332-1336

② 角野為耶、荒井ひろみ、中川裕志、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、情報処理学会論文誌、査読有、Vol.56,No.12, pp.2244-2252

〔学会発表〕（計 ５ 件）

① 角野為耶、中川裕志、 滞在場所のk-匿名化法、第 28 回人工知能学会全国大会、2B4-0S-15b-1，松山市 ひめぎんホール，2014 年 5 月 13 日

② 中川裕志、匿名化の実社会での利用へ向けての技術課題、第 28 回人工知能学会全国大会、2B4-0S-15a-4、ひめぎんホール(愛媛県･松山市)、2014 年 5 月 13 日

③ 中川裕志、ビッグデータ利用における個人データ保護における課題、情報処理学会、第 64 電子化知的財産・社会基盤研究発表会(EIP 研究会) Vol.2014-EIP-64、No.11.2014 年 5 月 15 日

④ 角野 為耶、中川 裕志、k-匿名化が誘発する濡れ衣を軽減するデータベース分割再構成法、The 32nd Symposium on Cryptgraphy and Information Security、リーガロイヤルホテル小倉(福岡県･北九州市)、2015 年 1 月 21 日

⑤ 中川裕志、匿名化の技術的俯瞰、 第 35 回医療情報学連合大会　シンポジウム 3 ICT による自己情報コントロールとプライバシー保護、 4-A-2-2、 沖縄コンベンションセンター（沖縄県･宜野湾市）、 2015 年 11 月 4 日

〔図書〕（計　1　件）
① 中川裕志、勁草書房、プライバシー保護入門 法制度と数理的基礎、第 6 章、2015、pp.135-147

〔産業財産権〕
○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
① SlideShare のアップロードコンテンツ k-匿名化と濡れ衣.2014.4.1 時点での閲覧数:5315. http://www.slideshare.net/hirsoshnakagawa3/k-31921914
②SlideShare のアップロードコンテンツ k-匿名化が誘発する濡れ衣を軽減する方法. 2014.4.1 時点での閲覧数:1576. http://www.slideshare.net/hirsoshnakagawa3/k-34727090

６．研究組織
(1)研究代表者
中川　裕志（NAKAGAWA,　Hiroshi）
東京大学・情報基盤センター・教授
研究者番号：２０１３４８９３
(2)研究分担者
佐藤　一誠（SATO,　Issei）
東京大学・新領域創成科学研究科・講師
研究者番号：９０６１０１５５