2B5-OS-15b-1

滞在場所の k-匿名化法

A method of k-anonynimity for protecting location privacy

角野為耶 *1　中川裕志 *2
Nasuka Sumino　Hiroshi Nakagawa

*1東京大学大学院学際情報学府　*2東京大学情報基盤センター
The University of Tokyo　The University of Tokyo

In the field of privacy preserving data mining, k-anonymity is a representative model for protecting privacy. But,existing methods of k-anonymity have a problem that a person is misleadingly suspected as a bad guy due to the information which actually has nothing to do with him/her.In order to prevent an occurrence of misleading, we have to pay attention to sensitive area when we generate anonymizing spatial region.In this paper, we propose an spatial anonymizing method with the use of constrained k-means that is aware of sensitive area , which can prevent an occurrence of misleading.

1. はじめに

パーソナルデータは個人 ID（氏名）、擬似 ID（住所、年齢、性別、国籍など）、センシティブ情報（宗教、病名、位置情報、収入などの他人に知られたくない情報）から構成される。データの匿名化とはパーソナルデータが持つこれらの情報から個人を特定出来ないようにデータを変換する操作であり、代表的な匿名化手法として k-匿名化が提案されている。k-匿名化とは、データが持つ個人 ID を消去ないし仮名化した上で擬似 ID の情報の一部を消去あるいは精度を落とし、全く同じ擬似 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匿名化手法である [Sweeney2002]。この操作を施すことにより、データベース内で個人を一意に特定することは不可能となる。表 2 に、表 1 を 3-匿名化したデータを示す。k-匿名化の中でも位置情報における匿名化の研究が進められている。通常の k 匿名化がデータテーブル内の全ての擬似 ID が同一となることで k-匿名化と判定するのに対して、位置情報の k-匿名化は位置情報の精度のみを落とし、精度を落とした領域内に k 人の人間が含まれればそれで k-匿名化が成されたものとみなす。このように精度を落として生成された、k 人を含むような領域を匿名化領域と呼ぶ。本論文では位置情報の k-匿名化について議論し、既存の匿名化手法が抱える濡れ衣という問題を示した上で濡れ衣を軽減する新たな匿名化手法を提案する。

表 1: 滞在場所のデータベース例

名前	年齢	性別	住所	N 月 M 日 P 時の所在
一郎	35	男	文京区本郷 WW	消費者金融
次郎	30	男	文京区湯島 YY	T 大学
三子	33	女	文京区弥生 ZZ	T 大学

表 2: 3-匿名化の例

名前	年齢	性別	住所	N 月 M 日 P 時の所在
A	30 代	*	文京区	消費者金融
B	30 代	*	文京区	T 大学
C	30 代	*	文京区	T 大学

連絡先: 角野為耶，東京大学，東京都文京区本郷 7-3-1，03-5841-2729, qq136316@iii.u-tokyo.ac.jp

2. k-匿名化が誘発する濡れ衣

k-匿名化が誘発する問題として、匿名化を行ったことによる濡れ衣の発生が考えられる。例を挙げると、表 2 において次郎、三子の 2 人は元々は消費者金融以外の場所に滞在していたが匿名化を施したことによって一郎と区別がつかなくなってしまったために消費者金融にいたことを疑われる恐れがある。このような、k-匿名化を行ったことによって身に覚えのない疑いをかけられる現象を k-匿名化が誘発する濡れ衣と呼ぶ [中川 2013]。濡れ衣の発生を防ぎ、且つ k-匿名化を行う方法としては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。例えば、消費者金融店舗に出入りしたのが 20 名中 2 名であれば、本当に消費者金融に滞在したユーザーは全体の 10 分の 1 の割合であり、他の 18 人を疑う労力は骨折りであるという心理が働くと考えられる。[中川 2013] 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。第二の方法は、消費者金融のような濡れ衣を発生させる恐れがある場所に滞在したユーザーを別々の領域に割り当てて、一つの匿名化領域に存在する濡れ衣を発生させうるユーザーの割合を小さくすることである。濡れ衣を発生させる恐れがある場所に滞在したユーザーを 2 名としてこれらのユーザーを k=10 の２つの匿名化領域に割り当てれば、先ほどと同様に匿名化領域に含まれる消費者金融に滞在したユーザーの割合を 10 分の 1 に抑えることができ、且つ k を小さく保てるのでデータの損失を抑えることが出来る。本論文では、第二の方法の考え方に基づいて位置情報の k-匿名化法を提案する。

3. 位置情報の k-匿名化に関する研究

位置情報の k-匿名化を実現する手法としては、Interval Cloak[Gluster2003]、Casper[Mokbel2006] などが提案されている。Interval Cloak は二次元空間を再帰的に四分割していき、分割された空間に位置情報を格納する。最も大きな矩形を最上位のノードとし、最も小さな矩形を最下位のノードとする。位置情報を記録する際には最も小さな矩形のノードから始めて領域内に含まれる人数が k 人以上になるまでノードを移動し、k 人以上になったノードの矩形を位置情報として記録する。この手法には、現在のノードに k 人以上の人間が含まれていない場合は 4 倍の大きさを持つ上位のノードに移動するためオーバーヘッドが発生してしまうという問題点がある。

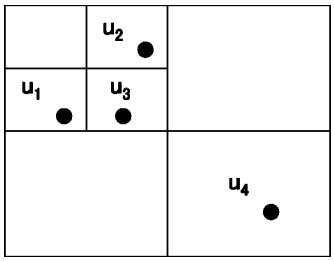

図 1 Interval Cloak

Casper は Interval Cloak と同様にと同様に二次元空間を再帰的に四分割し、矩形を位置情報とする匿名化手法である。現在のノードに k 人以上存在しない場合に隣接する同レベルの矩形を探索し、k 人以上の人数を含んでいれば隣接する矩形を足し合わせた矩形を位置情報として、上位ノードへの探索を打ち切るという点が Interval Cloak とは異なる。この操作によって、Interval Cloak が抱えていたオーバーヘッドの問題を改善している。これら全ての手法に共通して言えることは、トップダウン的に領域を分割していくため匿名化領域生成の際に濡れ衣の発生を考慮していないという点にある。本論文では、消費者金融のような濡れ衣を発生させうる場所をセンシティブな施設として定義し、センシティブな施設に着目したボトムアップ的な匿名化領域に生成手法を提案し、濡れ衣の発生を防ぐことを試みる。

4. 濡れ衣を軽減する k-匿名化

k-匿名化において濡れ衣が発生する確率は、個人の主観確率に依存し、匿名化領域内に存在するセンシティブなデータ点の割合に依存する。[中川 2013]。主観確率は匿名化領域内に存在するセンシティブなデータ点の割合の増加に対して、図 2 のような曲線を描きながら急激に増加する。従って濡れ衣の発生を防ぐには、一つの匿名化領域に存在するセンシティブなデータ点の割合を図 2 の β より低い値に保つ必要がある。

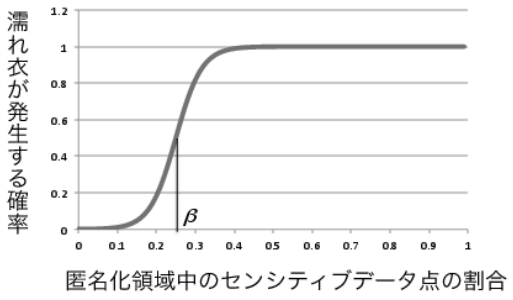

図 2 濡れ衣が発生する確率

本論文では、濡れ衣を軽減する為にセンシティブな施設に滞在したデータ点を別々の匿名化領域に割り当てることで、匿名化領域内のセンシティブなデータ点の割合を小さくすることを目指し、これを実現する為に制約付き k-means を用いたデータ点のクラスタリングを行う。本論文で提案する手法はデータの初期化、クラスタリング、再構成の 3 つに分けることが出来る。

4.1 表記

本論文で用いる表記を表 3 にまとめた。

表 3: 表記

表記	意味
N	空間全体中のデータ点の数
k	匿名化する人数
P=$\{p_i \mid i = 1, 2, \ldots, N\}$	空間全体中のデータ点の数
K	クラスタ数
C=$\{c_i \mid i = 1, \ldots, N\}$	クラスタ
$center(C_i)$	C_iの中心点
DS=$\{DS_i \mid i = 1, \ldots, m\}$	領域全体中のセンシティブ施設
m	センシティブな施設の数
$DS_i = \{s_{i,1}, \ldots, s_{i,\mid DS_i}\}$	センシティブ施設
$S = \{ps_i, \ldots, ps_l\}$	センシティブなデータ点
l	センシティブなデータ点の数
$s_{i,j}(x)$	$s_{i,j}$の x 座標
$s_{i,j}(y)$	$s_{i,j}$の y 座標
$NC(p_i)$	点 p_iから最も近いクラスタ
$NP(C_i)$	C_iから最も近いデータ点
$dist(a, b)$	点 a と点 b 間の距離
$far(C_i)$	C_i中最も中心から遠いデータ点
L	再構成に関わるクラスタ数
I	再構成で減らすクラスタ数
$RC = \{RC_i \mid i = 1, \ldots, L + 1 - I\}$	再構成されたクラスタ

4.2 初期化

初期化プロセスでは、まずセンシティブな施設に滞在したデータ点の位置情報を修正する。空間全体に存在するセンシティブ施設について、それぞれの施設から最も近いセンシティブ施設に滞在していないデータ点までの距離 Δ_i を求める。Δ_i をセンシティブなデータ点の位置の修正幅として、データ点をセンシティブ施設を中心とした半径 Δ_i の円上に位置を移動する。図 2 の例では、半径 Δ_i の円上に 3 つのセンシティブなデータ点を $\frac{2\pi}{3}$ の間隔で設置する。このような操作を行うことで、小さな領域に密集していたセンシティブなデータ点を別々の匿名化領域に格納することが可能となる。次に、初期クラスタの中心を決定する。初期クラスタの選択は 2 つのプロセスに分類される。センシティブなデータ点を別々の匿名化領域に格納するために、始めにセンシティブなデータ点それぞれをクラスタの中心として決定する。次に、残りのクラスタの中心を k-means++[Arthur2006] に従って選択する。ただしこのときセンシティブなデータ点の数は全体のクラスタ数 K よりも小さいものとする。この条件を満たさない場合はデータ点全体におけるセンシティブなデータ点の割合が非常に高い場合であり、そのような場合は濡れ衣を軽減しようがないため今回は考慮しない。初期クラスタの中心を選択した後、Algorithm3 に記載したクラスタ割当アルゴリズムによって各データ点をクラスタに割り当てて初期クラスタを決定する。図 3 の例では、黒く塗りつぶされたデータ点がアルゴリズムによって選ばれた初期クラスタの重心となる。

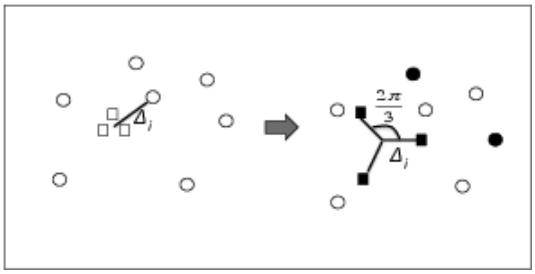

図 3 センシティブなデータとクラスタの初期化

4.3 クラスタリングアルゴリズム

本論文では制約付き k-means 法を用いてデータ点のクラスタリングを行い、各クラスタを匿名化領域とすることで位置情報の k-匿名化を実現する。クラスタリングプロセスはクラスタの中心更新プロセスとクラスタ割当プロセスから構成される。クラスタの中心更新プロセスでは、k-means 法と同様の手順でデータ点が割り当てられた各クラスタの中心を求める。

Algorithm 1 初期化アルゴリズム

$K \leftarrow \lfloor \frac{N}{k} \rfloor$
$\Delta_i \leftarrow DS_i$の中心から最も近いセンシティブでないデータ点までの距離
for $i = 1$ to m **do**
  for $j = 1$ to $|DS_i|$ **do**
    $s_{i,j}(x) \leftarrow center_x(DS_i) + \Delta_i cos(2\pi \frac{j}{|DS_i|})$
    $s_{i,j}(y) \leftarrow center_y(DS_i) + \Delta_i sin(2\pi \frac{j}{|DS_i|})$
  end for
end for
$h \leftarrow 0$
for $i = 1$ to m **do**
  for $j = 1$ to $|DS_i|$ **do**
    $center(C_h) \leftarrow s_{i,j}$
    $h \leftarrow h + 1$
  end for
end for
for $i = h$ to K **do**
  $k - means + +$に基づいて中心を選択
end for
for $i = 1$ to N **do**
  $AssignCluster(p_i, C)$
end for

クラスタ割当プロセスでは、一つの匿名化領域に属するデータ点の数を k 個とし、センシティブなデータ点が同じクラスタに属さないという制約を踏まえてデータ点をクラスタに割り当てて行く。全データ点について、各データ点と最も中心が近いクラスタにデータ点を割り当てることを試みる。このとき、割り当てようとしているデータ点がセンシティブなデータ点であり、且つ割り当てようとしているクラスタに既にセンシティブなデータ点が入っている場合は割り当てようとしているデータ点を次に中心が近いクラスタに割り当てる。データ点を割り当てようとしているクラスタに既に k 個の要素が含まれている場合は、そのクラスタの中で最も中心から距離が遠いデータ点と中心との距離、及び割り当てようとしているデータ点と中心との距離を比較して距離が近いデータ点をクラスタに割り当てる。このとき、クラスタに割り当てられなかったデータ点は次に中心の距離が近いクラスタに同様のプロセスで割り当てることを試みる。

Algorithm 2 クラスタリングアルゴリズム

while $C_{new} \neq C_{prev}$ **do**
  for $i = 1$ to K **do**
    $set\,center(C_i)$
  end for
  $AssignCluster(p_i, C)$
end while

4.4 クラスタ再構成アルゴリズム

クラスタ割当プロセスでは、一つのクラスタ当たりの要素数に制約をかけているため、図 3 の黒く塗りつぶされたデータ点で構成されるような歪な形のクラスタが出来てしまう恐れがあり、このような歪な形のクラスタを再構成する必要がある。そこで、Algorithm1,2 の終了後にクラスタの再構成を行う。再構成のアルゴリズムを Algorithm4 に示す。再構成プロセスではまずクラスタ内の各データ点と中心までの距離の二乗和を算出し、二乗和が大きいクラスタから順に再構成の対象としていく。次に、対象となったクラスタの周囲 L 個のクラスタを取り込み、再構成対象のクラスタと取り込んだクラスタのデータ

Algorithm 3 クラスタ割当アルゴリズム

$AssignCluster(p_i, C)$:
  if $p_i \in S$ **and** $(S \cap NC(p_i)) \neq \phi$**{**
    $AssignCluster(p_i, C \backslash NC(p_i))$**}**
  else if$|NC(p_i)| < k$**{**
    $NC(p_i) \leftarrow NC(p_i) \cup p_i$**}**
    else {
      if p_iの方が重心まで近い場合
        $NC(p_i) \leftarrow (NC(p_i) \backslash far(NC(p_i)) \cup p_i$
        $AssignCluster(far(NC(p_i)), C \backslash NC(p_i))$
      else
        AssignCluster$(p_i, C \backslash NC(p_i))$**}**
    }

点から新たに L+1-I 個のクラスタを形成する。なお、L 及び I は予め定めておくパラメータである。クラスタの形成に当っては Algorithm1 の初期化と同様の手順で初期クラスタの中心を決定し、次に 2 つのステップを経てデータ点をクラスタに割り当てて行く。第一のステップでは $ReAssign$ を用いて再構成クラスタをベースにデータ点を割り当てていく。$ReAssign$ は同じクラスタ内にセンシティブなデータ点が複数存在しないようにデータ点を割り当てるメソッドであり、各再構成クラスタの要素数が k 個になるまでこれを用いてデータ点を割り当てる。第二のステップでは、第一のステップで割り当てられなかった残りのデータ点をそれぞれ最も中心が近いクラスタに割り当てて行く。これらの 2 段階のプロセスを経てデータ点を割り当てることによって、クラスタ内の距離の分散が大きいクラスタの発生を防ぐ。クラスタ内分散が大きい者から順に再構成を行っていき、再構成後と再構成前のクラスタの分散を比較して再構成後の方が再構成前よりも分散が大きくなるまで再構成を行う。

Algorithm 4 クラスタ再構成アルゴリズム

クラスタ内分散の大きさ順に全クラスタをソート
3〜16 の処理を再構成によって分散が増加するまで繰り返す
for $i = 1$ to K **do**
  $T \leftarrow C_i$とその周囲 L 個のクラスタ内のデータ点
  for $i = 1$ to $L + 1 - I$ **do**
    $RC_i \leftarrow T$ から初期クラスタ中心点を初期化プロセスと同様の手順で選択し、各再構成クラスタに格納する
  end for
  9〜11 の処理を k 回繰り返す
  for $i = 1$ to $L + 1 - I$ **do**
    $ReAssign(T, RC_i)$
  end for
  for $i = 1$ to $k * I$ **do**
    残ったデータ点を最も近いクラスタに割り当てる
  end for
  C から再構成に関わったクラスタを取り除く
end for

$ReAssign(T, RC_i)$:
  if $NP(RC_i) \in S$ **and** $(S \cap RC_i) \neq \phi$**{**
    $ReAssign(T \backslash NP(RC_i), RC_i)$**}**
  else {
    $RC_i \leftarrow RC_i \cup NP(RC_j)$
    $T \leftarrow T \backslash NP(RC_j)$
  }

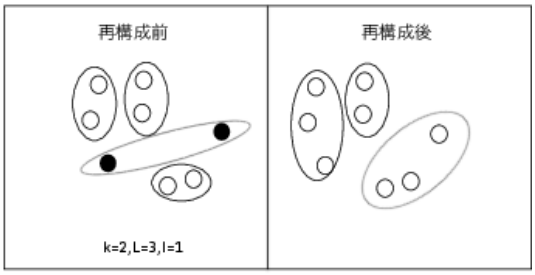

図 4 クラスタの再構成

5. 評価実験

二次元平面上にランダムにマッピングした人工のデータ点を用いて実験を行った。二次元平面の大きさを 100 × 100、データ点の数を 400、センシティブな施設を 2 棟、センシティブなデータ点を 5 つとした。再構成に関わるパラメータは L=4,I=1 とした。これらの実験データに対して Interval Cloak、Casper、提案手法を用いて匿名化を施し、評価指標を用いて匿名化の精度を評価した。評価指標には、濡れ衣を被る確率と情報損失の２つの指標を用いた。

5.1 濡れ衣による被害

先行研究 [中川 2013] を参考に、濡れ衣を評価する指標としてシグモイド関数を用いて既存手法と提案手法を評価した。

$$\text{濡れ衣の評価関数 } f(x) = \frac{1}{1-e^{\alpha(x-\beta)}}$$

今回の実験では α=30 として、β=0.15,0.2,0.25 の場合で実験を行い、濡れ衣を評価した。

表 4: 濡れ衣が発生する確率 k=5

	$\beta = 0.15$	$\beta = 0.20$	$\beta = 0.25$
IC	0.858	0.653	0.650
Casper	1.000	0.999	0.999
提案手法	0.743	0.431	0.154

表 5: 濡れ衣が発生する確率 k=10

	$\beta = 0.15$	$\beta = 0.20$	$\beta = 0.25$
IC	3.04e-2	4.94e-3	3.36e-5
Casper	7.97e-1	4.82e-1	6.40e-2
提案手法	6.67e-3	4.53e-5	3.05e-7

実験の結果、提案手法は全ての β において既存手法と比較すると濡れ衣を軽減出来ることが分かった。センシティブなデータ点が一つの匿名化領域に集中しているため、Casper は濡れ衣が発生する確率が非常に高くなっている。一方、IC は Casper と同様にセンシティブなデータ点が集中してはいるが、Casper と比べて一つの領域当たりに含まれるデータ点の数が多いため、Casper と比較すると濡れ衣が発生しにくい。k=5 の場合において提案手法は β の値によって確率が大きく変動しているが、これは β の値を評価関数が急激に上昇する領域に設定しているからだと考えられる。また、k=10 の場合においては一つの匿名化領域当たりに含まれるセンシティブなデータ点の割合が β の値と比べて小さくなっているため、どの β の場合でも濡れ衣が発生する確率が非常に低くなっていると考えられる。k=5 の場合と比較すると k=10 の場合ではどの手法でも濡れ衣が発生する確率は低くなっているが、k=5 の場合と比較して k=10 の場合は情報損失が非常に大きくなっており、情報損失を考慮しながら濡れ衣を軽減するには提案手法が最も優れた手法であると言える。

5.2 情報損失

情報損失を評価する指標として、クラスタ内の各データ点とクラスタ中心までの距離の二乗を用いる。実験の結果、提案手法は IC と比較すると情報損失が小さく、Casper と比較すると情報損失が大きいという結果が出た。提案手法ではセンシティブなデータ点が同じクラスタに存在してはならないという制約をつけてクラスタリングを行っており、制約によって歪な形のクラスタが出来上がっていることが原因だと考えられる。

表 6: 情報損失

	IC	Casper	提案手法
k=5	24.60	13.75	19.34
k=10	88.35	43.75	45.37

6. おわりに

本論文では、位置情報における濡れ衣を軽減する k-匿名化手法を提案した。提案手法は 3 段階から成り、第 1 段階ではデータ点とクラスタを初期化し、第 2 段階では制約付き k-means を用いて匿名化領域を生成し、第 3 段階ではクラスタの大きさが非常に大きいクラスタについて周囲のクラスタと合併させてクラスタを再構成した。人工データを用いた評価実験により、提案手法を用いて濡れ衣を軽減した匿名化を実現出来ることが確認出来た。

参考文献

[Sweeney02] Sweeney L. k-anonymity: A model for protecting privacy.*International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*,10.05:557-570, 2002

[Arthur07] Arthur D and Vassilvitskii S. k-means++: The advantages of careful seeding.*In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*,2007.

[Grusteser03] Gruteser,Marco,and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking.*Proceedings of the 1st international conference on Mobile systems, applications and services* ,2003.

[Mokbel03] Mokbel,Mohamed F, Chi-Yin Chow, and Walid G. Aref. The new Casper: query processing for location services without compromising privacy.*Proceedings of the 32nd international conference on Very large data bases*,2006.

[Nakagawa13] 中川裕志, 角野為耶. 滞在場所の k-匿名化と濡れ衣. 情報処理学会研究報告. *EIP,/*電子化知的財産・社会基盤*/* 62:1-6,2013.