RA-001

弱 ℓ-多様性問題を解く 3 近似アルゴリズム
3-approximate algorithm for weak ℓ-diversity

小池 優 † 赤木 俊裕 ‡ 宮田洋行 § 中野 眞一 §
Yu Koike Toshihiro Akagi Hiroyuki Miyata Shin-ichi Nakano

1. あらまし

大量のデータを分析して有益な知識を得ることに関する多くの研究がある．大量のデータを公開することが多くの分野で盛んになりつつある．

一般にできるだけオリジナルデータを公開したい．しかし，個人情報保護の観点から，データに対応する個人を特定されないようにしたい．

個人を特定されないように，公開するデータから個人情報を削除することが広く実施されている．しかし，これだけではプライバシー保護として不十分であることが最近の研究で明らかになっている．たとえば，公開するデータから，氏名や電話番号など，個人をほぼ特定できてしまうデータを削除しても，他の公開データを利用することにより,（例えば，郵便番号，性別，誕生日のデータから,）多くの個人がほぼ特定できてしまうことが報告されている [1]．このため，個人を特定されないように公開するデータを加工し，匿名化することが必要となる．

匿名化の基本的な方法は一般化である．例えば，元のデータでは年齢が 22 歳となっているものを一般化後のデータでは 20-29 歳とするように数値の値を区間に変更する．データを一般化することで，プライバシーを保護しやすくなるが，データは曖昧になり，有用性は低下してしまう．つまり，匿名化において，データの有用性と匿名性はトレードオフにあるといえる．また，公開するデータの中で，重要なもの（病歴など）は一般化をせず，そのまま公開することが望ましい．

このように，データの有用性をできるだけ損なうことなく，匿名化する方法が必要とされている．本研究では，このような匿名化の主要なモデルの一つである ℓ-多様性 [2, 6, 7] に関する問題を解く近似アルゴリズムを与える．

2. 準備

図 1 の表のデータを匿名化することを考えよう．各行をレコードとよぶ．各列を属性とよぶ．表の属性を，次に示す 3 つの種類に分類する．

- 識別子
  住所，氏名，電話番号のように，レコードに対応する個人をデータからほぼ特定できてしまう属性を識別子とよぶ．これらは，公開するデータから削除することが多い．
- 準識別子 (Quasi-Identifier, QI)
  郵便番号，性別，誕生日のように，他の公開データを利用することにより，レコードに対応する個人をほぼ特定できるかもしれない属性の集合を準識別子という．（一般化などの匿名化処理を行う．）
- センシティブ属性 (Sensitive Attribute, SA)
  病歴など，レコード保持者が秘密にしたい属性をセンシティブ属性という．

	氏名	電話番号	住所	年齢	体重	病歴
A		090-7234-****	群馬県桐生市***	22	52	胃炎
B		090-4654-****	東京都中央区***	48	72	糖尿病
C		080-7981-****	埼玉県川越市***	25	65	肺炎
D		090-5987-****	東京都渋谷区***	24	68	インフルエンザ
E		080-2231-****	群馬県みどり市***	59	66	糖尿病
F		090-9974-****	埼玉県浦和市***	51	60	糖尿病
G		090-16**	埼玉県大宮市***	33	75	心臓病
H		090-1789-****	東京都練馬区***	78	64	インフルエンザ
I		080-6589-****	栃木県栃木市***	78	57	肺炎
J		080-6731-****	群馬県前橋市***	74	48	アルツハイマー
K		090-3771-****	千葉県千葉市***	65	50	肺炎
L		090-1674-****	神奈川県横浜市***	64	53	胃炎
M		080-9632-****	茨城県水戸市***	74	59	アルツハイマー

削除

図 1: データの例

匿名性の指標にはいくつものモデルがある．主要な 3 つのモデルを紹介する．

2.1 k-anonymity[3]（k-匿名性）

例えば，年齢が 22 歳のデータがひとつしかないとき，名前などの識別子を削除しても，どのデータが 22 歳の人物に対応するかが分かってしまう．データを一般化することにより，特定の準識別子データが必ずいくつかあるようにすると，このようなことは防ぐことができる．

各レコードについて，準識別子データが同一なレコードが他に $k-1$ 個以上あるとき，k-匿名性があるという．例を表 1 に示す．すなわち，準識別子データを一般化することにより，準識別子データが同じレコードが必ず k 個以上あることを保証し，特定の個人のレコードがこのうちのどれに当たるかがわからないようにする．

2.2 ℓ-diversity[2]（ℓ-多様性）

k-匿名性だけでは解決できない次のような問題がある．ある人物の準識別子データがわかっているとき，同じ準識別子データをもつ k 個以上のレコードのうち，その人物のレコードがどのレコードに対応するかは k-匿名性により特定できない．しかし，もしそれらのレコードの集合に病歴データが 1 種類しかないならば，その人物のセンシティブ属性データである病歴が特定されてしまう．

†群馬大学理工学府電子情報・数理教育プログラム
‡群馬大学理工学府電子情報・数理領域
§群馬大学理工学府

Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 1: k-匿名性 (k=3) の例

氏名	年齢	体重	病歴
A	20-39	50-69	胃炎
C	20-39	50-69	肺炎
D	20-39	50-69	インフルエンザ
G	20-39	50-69	心臓病
B	40-59	60-79	糖尿病
F	40-59	60-79	糖尿病
E	40-59	60-79	糖尿病
H	60-79	50-69	インフルエンザ
K	60-79	50-69	肺炎
L	60-79	50-69	胃炎
J	70-89	40-59	アルツハイマー
I	70-89	40-59	肺炎
M	70-89	40-59	アルツハイマー

例えば，表 1 では，ある人物が B,E,F のいずれかのレコードに対応するとき，この人物がどのレコードに対応するかはわからないが，病歴はどのレコードも糖尿病であるため（表 2 参照），この人物の病歴は糖尿病と特定されてしまう．

表 2: 病歴の偏り

氏名	年齢	体重	病歴
B	40-59	60-79	糖尿病
F	40-59	60-79	糖尿病
E	40-59	60-79	糖尿病

準識別子データが同一なレコードの集合において，センシティブ属性の特定のデータの出現確率がいずれも $1/\ell$ 以下であるとき ℓ-多様性があるという．このとき，上に説明した k-匿名性の問題を防ぐことができる．このように，ℓ-多様性は，k-匿名性と比べ,（$k=\ell$ のとき,）よりプライバシーを保護できる匿名化である．ただし，データはより曖昧になる．

2.3 弱 ℓ-diversity[2]（弱 ℓ-多様性）

準識別子データが同一なレコードの集合において，センシティブ属性のデータが必ず ℓ 種類以上あるとき，弱 ℓ-多様性があるという．

3. 弱 ℓ-多様性問題

本研究では，データを一般化する代わりに代表値で表すことで匿名化することを考え，施設配置問題の枠組みで匿名化を考える [6, 7]．また，ここでは簡単のため，準識別子が 2 つで，かつ，どちらも数値データであるときを扱う．

施設配置問題とは，施設配置候補地の集合, 各施設の配置コスト, 顧客の集合等が与えられたとき, 指定されたコストを最小にするような施設の配置と各顧客の施設への割り当てを計算する問題である．

ここでは，各顧客が準識別子データに対応し，顧客を割り当てる施設が代表するデータに対応する．顧客と施設の距離が情報損失に相当する．

多くの施設配置問題は NP 困難であることが知られており, 多項式時間で解くことは現在のところほぼ不可能と思われる．

本研究で扱う弱 ℓ-多様性問題を定義しよう．表 3 および，図 2,3 参照．

表 3: 弱 ℓ-多様性 (ℓ=4) の例

氏名	年齢	体重	病歴
A	20-49	50-79	胃炎
C	20-49	50-79	肺炎
D	20-49	50-79	インフルエンザ
G	20-49	50-79	心臓病
B	20-49	50-79	糖尿病
E	40-79	40-69	糖尿病
H	40-79	40-69	インフルエンザ
K	40-79	40-69	肺炎
J	40-79	40-69	アルツハイマー
F	50-89	50-69	糖尿病
L	50-89	50-69	胃炎
I	50-89	50-69	肺炎
M	50-89	50-69	アルツハイマー

- 入力

  顧客の集合：$C = \{c_1, c_2, ..., c_n\}$
  （C は平面上の点の集合であり，各 c_i は顧客に対応し，各点の x 座標と y 座標が 2 つの準識別子データに対応する.）

  顧客の病歴：$q_1, q_2, ..., q_n$
  （平面上の各点は顧客の病歴に対応する.）

  施設配置候補地の集合：$F = \{f_1, f_2, ..., f_m\}$
  （平面上の点の集合であり，代表値データの集合に対応する.）

  施設に割り当てる最低限の病歴の種類：ℓ
  （施設に割り当てる最低限の病歴の種類を整数で指定する.）

  コスト：$d(c, f)$
  （各 $c \in C$ と各 $f \in F$ の距離であり，ここで，d は三角不等式を満たすとする.）

- 出力

  各 $f \in F'$ について $|\{q_i | A(c_i) = f\}| \geq \ell$ であるような，C から $F' \subset F$ への割り当て $A : C \to F'$ で，$\max_{c \in C} d(c, A(c))$ が最小のもの

$\{c | A(c) = f\}$ はデータ f に代表されるレコードの集合に相当し，$\{q_i | A(c_i) = f\}$ はデータ f に代表されるレコードの色（病歴）の集合に相当する．$\max_{c \in C} d(c, A(c))$ はデータの曖昧性であり，オリジナルデータと一般化データの距離の最大値である．これをコストとする．データの曖昧性は元のデータからの情報の損失を表すため，曖昧さを減らすことで情報の有用性を保つことができる．

Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

4. 提案するアルゴリズム

弱 ℓ-多様性問題は NP 困難であるため，最適解を多項式時間で計算することは非常に困難であると考えられる．そこで，近似アルゴリズムを設計する．これは，r-gathering 問題を解く Best-or-Rest[4, 5] アルゴリズムを改造したものである．

$\ell b(c_i, f_j)$ を以下のように定める．これは $c_i \in C$ を $f_j \in F$ に他の $\ell - 1$ 点と共に割り当てるときにかかるコストの下界である．次の 2 つの場合に分けて定義する．f_j に近い順に顧客を並べたときに現れる ℓ 色目の初めての顧客を c_k とする．

1. $d(c_i, f_j) \leq d(c_k, f_j)$ のとき，$\ell b(c_i, f_j) = d(c_k, f_j)$
2. $d(c_i, f_j) > d(c_k, f_j)$ のとき，$\ell b(c_i, f_j) = d(c_i, f_j)$

$\ell b(c_i)$ を次のように定める．これは，$c_i \in C$ をいずれかの $f \in F$ に割り当てるときにかかるコストの下界である．

$\ell b(c_i) = \min_{f_j \in F} \ell b(c_i, f_j)$

これを実現する施設 f_j を c_i の best facility とよび，この f_j から各色ごとに最も近い点を選び，これらの中から，f_j に近い順に ℓ 番目までの点，および，点 c_i が含まれていなければ c_i を追加して得られる点集合を c_i の partners とよぶことにする．c_i の partners は ℓ 点もしくは，$\ell + 1$ 点の集合である．これらには ℓ 色の点があることに注意しよう．

Algorithm 1 find weak ℓ-diversity

```
for 各顧客 c_i に対して do
    best facility と partners を計算する
end for
for 各顧客 c_i に対して do
    if c_i の best facility f には顧客が未割り当て，
    かつ，c_i のすべての partners はいずれの facility にも未割り当て then
        c_i と c_i の partners を f に割り当てる
    end if
end for
for まだ割り当てていない各顧客 c_i に 対して do
    顧客を割り当て済みの施設配置候補地のうち，c_i からの距離が一番小さなものに c_i を追加して割り当てる
end for
```

このアルゴリズムは次の 3 ステップからなる．

1. 各顧客に対して，best facility と partners を計算する．
2. 可能な限り各顧客を，その partners と共に best facility にグリーディーに割り当てる．
3. 2 で割り当てられなかった各顧客を，顧客を割り当て済みの最も近い施設配置候補地に，追加して割り当てる．

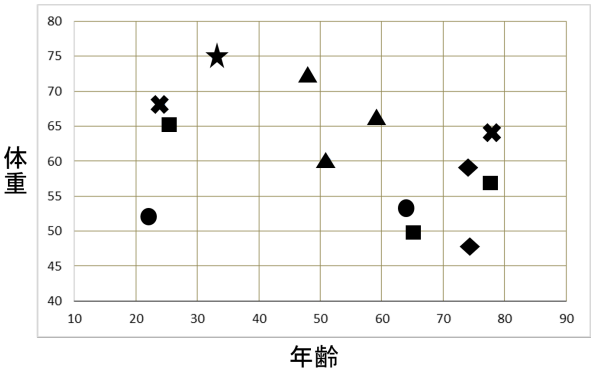

図 2: 顧客の配置例
（色の代わりに形で顧客の病歴を表している．）

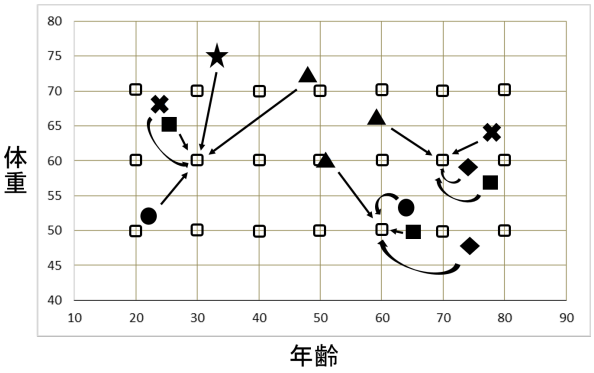

図 3: 顧客の施設配置候補地への割り当て例

4.1 正当性

ステップ 2 で顧客を best facility f に割り当てるとき，f には ℓ 個の全て異なる色の点が割り当てられる．ステップ 3 で顧客を追加して施設配置候補地に割り当てるとき，そこには，すでに ℓ 個の全て異なる色の点が割り当てられている．すなわち，いずれの場合も，割り当て先の各 $f \in F'$ について $|\{q_i | A(c_i) = f\}| \geq \ell$ が成立している．

4.2 近似比

ステップ 2 の割り当てのコストは，$\ell b(c_i)$ であり，これは最適値の下界である．ステップ 2 で割り当てられなかった各顧客は，顧客が割り当て済みの施設配置候補地に，ステップ 3 で追加して割り当てられる．この顧客 c_i がステップ 2 で割り当てられなかった理由は 2 つの場合がある．いずれの場合も最適解のコストの 3 倍以下のコストしかかからない．

（1）c_i の partners のいずれかがいずれかの施設配置候補地に割り当て済みのとき．partners 中の割り当て済みの顧客を c_j とする．c_j は c_k の partners 中の 1 点として，ステップ 2 で c_k の best facility f_k に割り当てられたとする．c_i の best facility を f_i とする．このとき，$d(c_i, f_k) \leq d(c_i, f_i) + d(c_j, f_i) + d(c_j, f_k) \leq 2(\ell b(c_i)) + \ell b(c_k)$ である．図 4 参照．$\ell b(c_i)$ と $\ell b(c_k)$

Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

はいずれも最適値の下界であるため，ステップ 3 での c_i の割り当てコストは最適解の 3 倍以下である．

（2）c_i の best facility に顧客が割り当て済みのとき．c_i が他の顧客の partners 中の 1 点として割り当てられることなくステップ 2 が終了したとき，最後にステップ 3 において，c_i は，c_i の best facility もしくは，より近い facility に追加して割り当てられる．すなわち，割り当てコストは最適値の下界以下である．

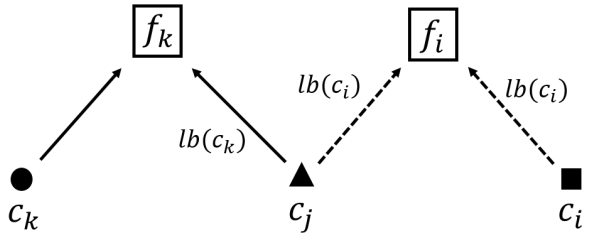

図 4: 近似比の説明図

4.3 計算時間

アルゴリズムの計算時間をステップごとに見積もろう．

1. まず，各施設配置候補地について最寄りの ℓ 色の顧客のリストを計算する．はじめに各施設から，各色ごとに最寄りの点を選ぶ．この中から，ℓ 番目に近い点を（線形時間のセレクションアルゴリズムにより）選び，そののちに，これより近い各色の点を選ぶ．以上は，ひとつの施設について $O(|C|)$ 時間かかり，これをすべての施設に対して行うと，$O(|C||F|)$ 時間かかる．次に，このリストを利用し，各顧客の best facility と partners を求める．各顧客ごとに $O(|F|)$ 時間かかり，全ての顧客では，$O(|C||F|)$ 時間かかる．

2. 各顧客 c_i を，c_i の best facility に割り当てられるかをチェックする．各 facility に checked flag を用意する．初めに off に初期化する．ステップ 2 では，まずこの flag を調べる．これが on のときは何もしない．一方，off のときは，flag を on にし，partners 中の各点が未割り当てかどうかをチェックし，もしそうなら best facility に割り当てる．これに $O(\ell)$ 時間かかり，全ての顧客では $O(\ell|F|)$ 時間かかる．各施設配置候補地ごとに，最寄りの ℓ 色の顧客は 1 度しかチェックしないことに注意しよう．

3. ステップ 2 で未割り当ての各顧客を，顧客が割り当て済みの最も近い施設配置候補地に割り当てる．顧客が割り当て済みの施設は 1 つ以上あることに注意しよう．各顧客について，$O(|F|)$ 時間かかり，全ての顧客では合計 $O(|C||F|)$ 時間かかる．

以上より，アルゴリズム find weak ℓ-diversity の計算時間は $O(|C||F|)$ 時間である．

5. まとめ

準識別子が 2 つの数値データの属性からなるとき，弱 ℓ-多様性問題を解く近似アルゴリズムを設計した．このアルゴリズムは，コストが最適解の 3 倍以内の近似解を求める．計算時間は $O(|C||F|)$ 時間である．

本手法は，準識別子が数値データであり，かつ，定数個の場合にも，（三角不等式を満たすならば,）最適解の 3 倍以内の近似解を求めるように一般化できる．

参考文献

[1] A. Froomkin, *“The Death of Privacy”,* Stanford Law Review, 52(5), pp.1461-1543 (2000).

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, *“L-diversity: Privacy Beyond K-anonymity”,* ACM Transactions on Knowledge Discovery from Data, 1(1), Article 3 (2007).

[3] L. Sweeney, *“k-Anonymity: A Model for Protecting Privacy”,* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp.557-570 (2002).

[4] A. Armon, *“On min-max r-gatherings”,* Theoretical Computer Science, 412, pp.573-582 (2011).

[5] T. Akagi, R. Arai, and S. Nakano, *“Faster min-max r-gatherings”,* IEICE TRANS. FUNDAMENTALS, to appear (2016).

[6] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, *“Fast Data Anonymization with Low Information Loss”,* VLDB ’07 Proceedings of the 33rd international conference on Very large data bases, pp.758-769 (2007).

[7] J. Li, K. Yi and Q. Zhang, *“Clustering with Diversity”,* Automata, Languages and Programming Volume 6198 of the series Lecture Notes in Computer Science, pp.188-200 (2010).

Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.