差分プライバシのプライバシー費用による一般化 k-匿名化の評価手法の提案

堀込　光 †

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室 †

1　はじめに

2017 年 5 月に施行された改正個人情報保護法では，本人の同意なく個人情報を第三者に提供できる匿名加工情報が導入された．しかし，k-匿名化 [2] における k などの明確な規定はなく，加工をどのレベルまで行うか不確かなことが課題であった．加工による有用性低下は定義しやすいが，その時の安全性は攻撃者の仮定に大きく依存して，精密に評価するのが難しいためである．例えば，Terrovitis らは再符号化の方程式を提案している [7] が，その安全性評価は不十分である．

そこで本研究では，攻撃者の計算能力向上によらない安全性の保証である差分プライバシに着目する．差分プライバシを満たすデータベースの作成手法，ラプラスメカニズム [1] や Radamized Response[6] では，プライバシ費用 ϵ によって決まるノイズを付加することにより差分プライバシを実現する．

これに対して，本研究では，ランダムサンプリングを行ったデータに一般化 k-匿名化を適用したデータからプライバシ費用 ϵ を計算することにより，差分プライバシの観点から k-匿名化の安全性を評価する手法を提案する．

2　従来技術

2.1　差分プライバシ

差分プライバシ [1] とは，2006 年に Dwork が提唱したプライバシの定義である．

定義 1. 任意の $S \in Range(A)$ と任意の隣接するデータセット D と D' についてランダムアルゴリズム A がある実数 ϵ について，

$$e^{-\epsilon} \leq \frac{Pr[A(D) = S]}{Pr[A(D') = S]} \leq e^{\epsilon}$$

を満たすとき，ϵ-差分プライバシを満たすという．

この時，プライバシ費用である ϵ が小さいほどデータベース D，D' の区別が難しくなり，安全性は高くなる．差分プライバシでは，任意の背景知識を持つ攻撃者や未知の攻撃者に対して安全性を保証する．

2.2　k-匿名

k-匿名 [2] は，2002 年に Sweeney により提唱されたプライバシ保護技術である．

定義 2. 準識別子の集合を $QI = \{A_1, \ldots, A_d\}$ とする．任意のレコード $t \in D$ に対して，任意の準識別子の値の組 $(a_1, \ldots, a_d)$ を持つレコードが全ての QI について k 以上存在しているとき，データベース D は k-匿名性を満たす．

2.3　一般化

k-匿名化の手法の一つとして一般化がある．一般化とは，属性の要素をより広い意味合いを持つ要素に置き換える操作である．一般木の例を図 1 に示す．属性値が近い要素同士で階層化クラスタリングし，構成していく．一般木の階層の深さは最大で 3 である．Local-gov（地方公務員）や State-gov（州公務員）は Government（公務員）のように一般化される．

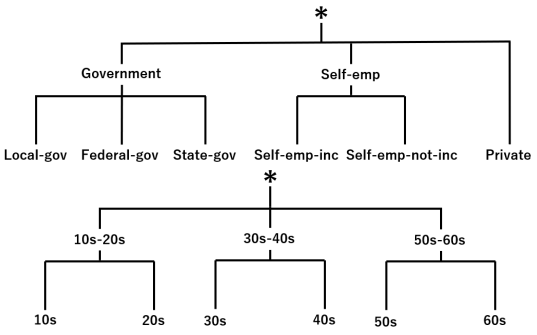

図 1　業種 (workclass) と年齢 (age) の一般木

2.4　NCP (Normalized Certainty　Penalty)

一般化による k-匿名化アルゴリズム，NCP (Normalized Certainty Penalty) 指標による細分化 [3] である．データベース D 内のレコード t_i の属性 Q_j の要素が $q_{j,u}$

†Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

であるとき，レコード t_i の属性 Q_j の NCP は，以下のように求められる．

$$NCP_{Q_j}(t_i) = \sum_{j=1} \frac{|q_{j,u}|}{|Q_j|}$$

この時，$|q|$ は，その要素 q が持つ葉ノードの数である．例えば，$t_i = [10s - 20s, Goverment]$ とすると，

$$NCP(t_i) = \frac{2}{6} + \frac{3}{6} = \frac{5}{6}$$

となる．詳細化可能な属性を詳細化した際に，D 内の全レコードの NCP_{Q_j} の和，

$$NCP_{Q_j} = \sum_{i=1} NCP_{Q_j}(t_i)$$

が最小となる属性を詳細化の対象とし，詳細化後の k-匿名を検討し，満たしている場合，詳細化を行う．また，満たしていない場合は，詳細化を行わず他の NCP が最小となる属性で詳細化を試みる．これを全ての属性が細分化できなくなるまで繰り返す．

Terrovitis らは大域的再符号化 [7] を提案している．この手法では，ある要素全てを同じ値に一般化する手法である．

$k = 5$ と $k = 10$ のときの Age(年齢) 属性の一般木の変化を図 2，図 3 に示す．ともに点線で詳細化の程度を示している．図 2 の k=5 のとき，age(年齢) 属性は，$10s - 20s$，$30s$，$40s$，$50s$，$60s$ に一般化される．$k = 10$ のとき，50 代と 60 代は $50s - 60s$ に一般化されいるのに対し，$k = 5$ では，一般化されず，k が小さいとき詳細なデータとなる．

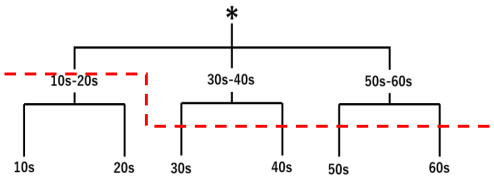

図 2　$k = 5$ の際の一般木

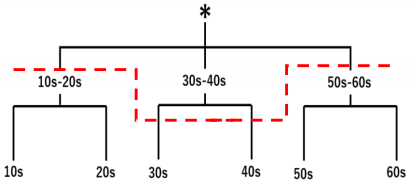

図 3　$k = 10$ の際の一般木

3　提案手法

3.1　サンプリングによる差分プライバシ

Kamalika らは差分プライバシの観点からランダムサンプリングの安全性について示している [4]．これを参考として，サンプリングを行ったデータベースに一般化 k-匿名化を行ったデータの差分プライバシについて次の命題 1 が証明できる．

命題 1. レコード t を n_t 個持つデータ D から，s_t 個を β のサンプリング率でランダムサンプリングしたデータ $A(D)$ は，

$$\epsilon = \log(1 - \beta) - \log(1 - \frac{s_t}{n_t})$$

について ϵ-差分プライバシを満たす．

Proof. アルゴリズム A をランダムサンプリングを行うアルゴリズム，サンプリング率を β とし，隣接するデータセットをデータベース D と任意のレコード t を抜いたデータベース D_{-t} とする．また，S をランダムサンプリングを行ったデータベースとし，データベース S 内のレコード t の数を s_t とする．今，D 内のレコード t の数を n_t とすると，D_{-t} 内のレコード t の数は $n_t - 1$ となる．

任意のレコード u を考えた場合，n_u 個のレコードから s_u 個サンプリングされる確率は，

$$Pr[n_u \in A(D) = s_u] = {}_{n_u}C_{s_u} \beta^{s_u} (1 - \beta)^{n_u - s_u}$$

$Pr[A(D) = S]$ は，すべてのレコードのサンプリング確率の積となるが，隣接するデータベース D と D_{-t} を考えると，レコード t 以外のレコード数は同じであるため，レコード t のみについてサンプリング確率を求めると，

$$\frac{Pr[A(D) = S]}{Pr[A(D_{-t}) = S]} = \frac{{}_{n_t}C_{s_t} \beta^{s_t} (1 - \beta)^{n_t - s_t}}{{}_{n_t - 1}C_{s_t} \beta^{s_t} (1 - \beta)^{n_t - s_t - 1}}$$

$$= \frac{n_t(1-\beta)}{n_t - s_t} = \frac{1}{1-\frac{s_t}{n_t}}(1-\beta) \leq e^{\epsilon}$$

両辺の対数を取り，命題を得る． □

ランダムサンプリングでは，ランダムサンプリングにより得られたデータ S と元のデータベース D のレコード数からプライバシ費用 ϵ を求めることができる．サンプリング後に k-匿名化を行う場合を次節で議論する．

3.2 サンプリング後の k-匿名化

レコード数 n のデータベース D からサンプリング率 β でランダムサンプリングを行い，k-匿名化するアルゴリズムを A とする．また，作成されたデータを $S = A(D)$ とする．

しかし，この時 $A(D) = S$ とはならない場合がある．例えば，$[age, workclass]$ の場合を考えてみる．age 属性の要素 $30s - 40s$ について詳細化を考えると，$[30s, Goverment]$, $[40s, Goverment]$，$[30s, Private]$, $[40s, Private]$，$[30s, Self]$，$[40s, Self]$ のすべてが k 以上のとき，age 属性の要素 [30s-40s] は詳細化してしまう．このような条件は，データ S 内のレコードが $2k \leq |[30s-40s, Goverment]|$, $2k \leq |[30s-40s, Private]|$, かつ，$2k \leq |[30s-40s, Self]|$ である．同様に，各要素についても全ての組み合わせが k 以上存在したデータとなる可能性を排除する必要がある．

Proof.

$$\frac{Pr[A(D) = S]}{Pr[A(D_{-t}) = S]} = \frac{{}_{d_t}C_{n_t}\beta^{n_t}(1-\beta)^{d_t-n_t}}{{}_{d_t-1}C_{n_t}\beta^{n_t}(1-\beta)^{d_t-n_t-1}} = \frac{d_t(1-\beta)}{d_t - n_t} \leq e^{\epsilon}$$

□

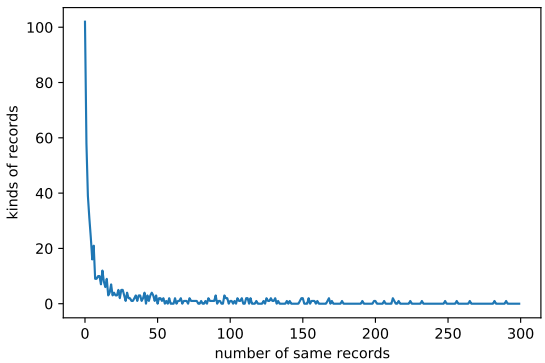

図 4　レコード数分布

4 実験

4.1 実験目的

k-匿名化による差分プライバシのプライバシ費用を評価する．

4.2 使用データ

本実験では Adult Data Set[5] を使用し，欠損値を含むレコードを削除した 45,223 レコードで実験を行った．使用する属性は，Age（年齢），Workclass（職業），Education（学歴），Income（所得）であり，Age 属性に関しては，一の位を切り捨てた数値を離散値として扱う．使用データのレコード分布を図 4 に示す．

4.3 実験結果

本実験では，匿名化指標 k とサンプリング率 β を変化させ一般化 k-匿名化のプライバシ費用を算出する．この時，5.2 節で述べた $A(D) = S$ とならない条件を無視する．匿名化指標を $3 \leq k \leq 15$，サンプリング率を $0.1 \leq \beta \leq 0.35$ についてプライバシ費用 ϵ 算出する．k を $k = 10$ に固定し，サンプリング率 β を変化させた際の ϵ の変化を図 5 に示す．図 6 には，サンプリング率 β を $\beta = 0.2$ に固定し，k を変化させた際の ϵ の変化を示す．プロットは 10 回ずつ S を出力した際の平均である．また，平均と標準偏差を表 1 に示す．

表 1　β と k における ϵ の平均と標準偏差

β	k	平均	標準偏差
0.1	10	0.7571	0.0134
0.15	10	0.8618	0.0203
0.2	10	1.0252	0.0331
0.25	10	1.1838	0.0321
0.3	10	1.4256	0.0362
0.35	10	1.9488	0.0612
0.2	3	1.5973	0.0296
0.2	5	1.0652	0.0262
0.2	15	0.4128	0.0191

4.4 考察

図 5 よりサンプリング率 β が増加するに伴い，ϵ も増加する．定数 $\alpha = 0$ の時，ϵ は，得られたデータ S と S 内の任意のレコードになり得る元のデータベース D

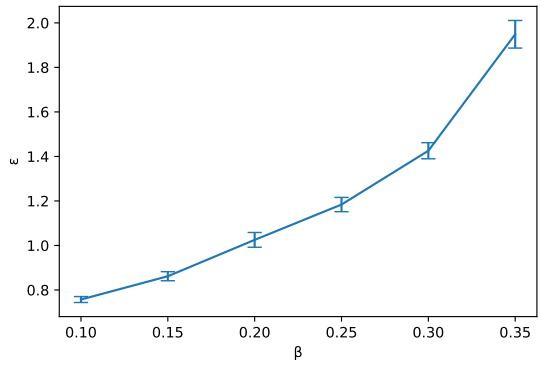

図 5　サンプリング率 β による ϵ の変化

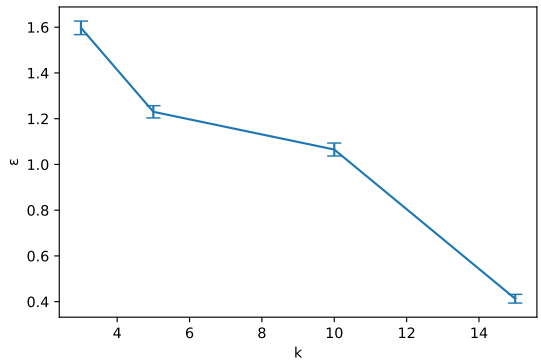

図 6　k-匿名の k による ϵ の変化

内のレコード数を比較することで算出できる．β を大きくすることで，d_i と n_i の値の差が小さくなり，分母である $d_i - n_i$ が小さくなる．一般化による k-匿名化では，$k = 10$，$\beta = 0.35$ で $\epsilon = 2.0$ 程度である．

また，図 6 より k が大きくなると ϵ が小さくなり，安全性が向上する．$k = 10$ 以上では，ϵ は 1 以下となり安全性は高いと言える．

5　おわりに

本実験では，Adult Data set を用いて差分プライバシの観点から一般化 k-匿名化を評価した．3.2 節で述べた $A(D) = S$ とならない組み合わせを定式化することが今後の課題である．また，他の差分プライバシ手法の有用性比較を行う必要がある．

参考文献

[1] C.Dwork，F.McSherry，K.Nissim，A.Smith，“Calibrating noise to sensitivity in private data analysis”，TCC, Vol. 3876, pp. 265–284, 2006.

[2] L.Sweeney，“k-anonymity: a model for protecting privacy”，International Journal on Uncertainty，Fuzziness and Knowledge-based Systems，10(5)，pp．557-570，2002.

[3] J.Xu，W.Wang，J.Pei，X.Wang，B.Shi，A.W.Fu，“Utility-based anonymization using local recoding”，Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining，pp．785-790，2006.

[4] Kamalika Chaudhuri，Nina Mishra，“When Random Sampling Preserves Privacy”，26th Annual International Cryptology Conference (CRYPTO)，pp．198-213，2006.

[5] UCI，“Adult Data Set”，(`https://archive.ics.uci.edu/ml/datasets/Adult/`，2020 年 6 月参照).

[6] C.Dwork，A.Roth，“The Algorithmic Foundation of Differential Privacy”，Foundations and Trends in Theoretical Computer Science，Vol．9，No．3-4，pp．211-407，2004.

[7] M.Terrovitis，N.Mamoulis，P.Kalnis，“Privacy-preserving Anonymization of Set-valued Data”，Proceedings of the VLDB Endowment，1(1)，pp．115-125，2008.