L-011

ヒストグラム密度推定に基づく匿名加工とその定量的評価

Data Anonymization via Histogram Density Estimation and Its Quantitative Evaluation

大島 朱音 [1]　　亀谷 由隆 [1]

Akane Oshima　　Yoshitaka Kameya

1 はじめに

個人情報や個人の特定に繋がるデータを利用，提供する際にはプライバシーに配慮しなければならない．プライバシーを保護するための技術として，データを加工して個人を特定しにくくする匿名加工が知られる [14]．平成 29 年 5 月に施行された改正個人情報保護法でも匿名加工に関する記述が追加された．プライバシーを保護しつつ，正確なデータ分析を行うためにどのように匿名加工するかが課題になっている．

年齢や性別等，間接的にその事例の人物を特定できる可能性がある属性は疑似識別子あるいは準識別子と呼ばれる．値の粒度が細かい数値属性が疑似識別子である場合，より特定されやすくなる．その場合，例えばその属性を区間分割し，各区間中の値をその区間内の値の平均で全て置き換える匿名加工が考えられる．また，各区間中の値をその区間の範囲内で一様分布に従って生成した値で置き換える方法も考えられる．前者はマイクロアグリゲーションと呼ばれる．また，後者はデータ合成と呼ばれる匿名加工の一例と言える．本論文ではこの 2 つの匿名加工を考える．

一方，匿名加工の基準としてよく知られるのが k-匿名性 [11] である．データベース中に同じ疑似識別子を持つ事例が k 件以上あるなら，そのデータベースは k-匿名性を満たすという．k-匿名性を満たせば，個人を特定する確率が $1/k$ 以下となるため，我々は k によって匿名性の度合を制御できる．前述の匿名加工について考えると，各区間に含まれる値（事例）の数が k 以上ならばマイクロアグリゲーションの結果は k-匿名性を満たす．また，データ合成においても k が大きければ加工後のデータから元の疑似識別子は推測されにくくなる．

一般に，個人の特定されにくさを意味する匿名性とデータ分析結果の正確性・信頼性を意味する有用性はトレードオフの関係にある．本論文では，有用性を保つ範囲で，なるべく高い匿名性を確保することを考える．そこで，以下では分割した各区間に含まれる事例数をその区間の「匿名度」と呼ぶ．そして，有用性を損なわない限り，この匿名度が大きくなる区間分割の方が望ましいと考える．一方，上述の k を最小限の匿名度を指定するパラメータと見なし，「最小の匿名度」と呼ぶことにする．

本論文では，単一の数値的な疑似識別子を含むデータに対して，有用性を保つ範囲で，なるべく高い匿名性を確保するように区間分割する手法を提案する．そして，この区間分割に基づき，マイクロアグリゲーションやデータ合成などの匿名加工が実施される．提案手法では，Kontokanen らのヒストグラム密度推定手法 [6] を用いながら，AIC [1] や BIC [9] に照らして最適な区間の分割点を求める．

これまでも単一の数値的な疑似識別子を含むデータに対する区間分割手法が提案されてきた [3, 12]．特に Hansen らの方法 [3] は，最小の匿名度 k を与えたときに区間内誤差二乗和 (the sum of within-group squared error, SSE) を最小にすることが保証されている区間分割法である．しかし，SSE は区間数が多いほど小さくなるため，この方法では結果的にほぼ全ての区間で事例数が k もしくはそれに近い数になる．それに対し，提案手法では，AIC や BIC の持つペナルティ項の働きにより，確率密度の観点から見て細かく分割する必要のない領域では広い区間分割を行い，より高い匿名性を確保することを目指す．

また，データを匿名加工した後に加工後のデータが本当に有用であるかを定量的に評価する手段が必要である．定量的な評価方法として，本研究では密度比推定に基づく Kullback-Leibler 情報量（以下，KL 情報量）の推定値 [8, 13] を用いる．著者らの知る限り，匿名加工データの有用性の評価において密度比推定に基づく KL 情報量の推定値を用いた例はない [2]．人工データと実データを用いた実験では，この評価方法に基づいて Hansen らの方法と提案手法の匿名性と有用性のバランスを評価する．

本論文は以下の構成をとる．はじめに 2 節では幾つかの記法を導入し，Hansen らの方法を紹介する．そして，3 節で本論文で提案する区間分割法を記述する．4 節では匿名加工データの有用性の評価手法について述べる．実験結果は 5 節で示す．最後に 6 節でまとめを行い，今後の課題を述べる．

2 準備

2.1 区間の分割

匿名加工の対象となっているサイズ n のデータを考える．そして，データ中に出現する相異なる疑似識別子を昇順で並べたものを $x_1, x_2, \ldots, x_m$ とおく $(1 \leq m \leq n)$．各 x_i の重複数を $w(x_i)$ と表記する．$\sum_{i=1}^{m} w(x_i) = n$ が成り立つ．ここで，x_e と x_{e+1} の中点 c_e で区間を分割することを考える $(1 \leq e \leq m-1,\ c_e = (x_e + x_{e+1})/2)$．また，$e$ 番目の分割点までの累積の事例数を n_e と書くと，$n_e = \sum_{i=1}^{e} w(x_i)$ が成り立つ．そして特別に $n_0 = 0,\ n_m = n$ と定める．

2.2 Hansen らの方法

このとき Hansen らの方法では，e 番目の分割点までの区間分割に対する SSE の部分スコアを

$$B_e = \min\left\{ \mathrm{SSE}_{0,e},\ \min_{1 \leq e' \leq e-1} \left\{ B_{e'} + \mathrm{SSE}_{e',e} \right\} \right\} \qquad (1)$$

と再帰的に定める $(1 \leq e \leq m)$．B_m が区間分割全体の SSE となる．ここで，$\mathrm{SSE}_{e',e}$ は e' 番目，e 番目の分割点間における誤差二乗和として下のように定義される．

$$\mathrm{SSE}_{e',e} = \sum_{i:e' \leq i \leq e} w(x_i)\,(x_i - \mathrm{Mean}_{e',e})^2 \qquad (2)$$

$$\mathrm{Mean}_{e',e} = \frac{1}{n_e - n_{e'}} \sum_{i:e' \leq i \leq e} w(x_i) x_i \qquad (3)$$

$\mathrm{Mean}_{e',e}$ は e' 番目，e 番目の分割点間の平均値である．そして，Hansen らの方法では SSE を最小にする区間分割を式 1 の再帰スコアの下で動的計画法に従って求める [3]．更に，最小の匿名度 k を設定されたとき，Hansen の方法では区間内の事例数が k 以上 $2k$ 未満になるように区間分割する [4]．

[1] 名城大学理工学部情報工学科

[2] 疑似識別子が離散値の場合はいくつか例がある [4, 5]．

[3] 提案論文 [3] では，区間内誤差二乗和を辺重みとした辺重み付き有向グラフをデータから構築し，その上で最短経路長問題を解くことで SSE を最小とする区間分割を見つけるが，ここではより単純な形で記述している．

[4] 本研究の比較実験では，疑似識別子の重複によって区間分割できない場合に限ってこの上限 $2k$ を破ることを許すように実装した．

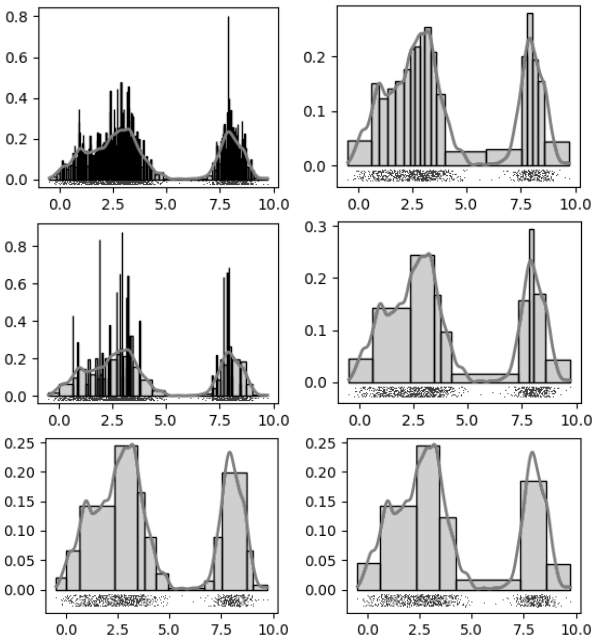

図 1: サイズ 1,000 の人工データに対する（上段）Hansen の方法の分割結果,（中段）AIC を用いた提案手法の分割結果,（下段）BIC を用いた提案手法の分割結果．左列は $k = 5$, 右列は $k = 50$ の場合．実曲線はカーネル密度推定をした結果．

例として，3 つの正規分布 $\mathcal{N}(1.0, 0.6^2)$, $\mathcal{N}(3.0, 0.8^2)$, $\mathcal{N}(8.0, 0.5^2)$ をそれぞれ混合比 0.2, 0.5, 0.3 で混合した混合正規分布から生成した 1,000 個のデータに対して Hansen らの方法を適用した分割結果を図 1 上段に示す（$\mathcal{N}(\mu, \sigma^2)$ は平均 μ, 分散 σ^2 の正規分布を表す）．左が最小の匿名度 k を 5 に設定した場合，右が k を 50 に設定した場合である．先述の通り，SSE は区間数が多いほど小さくなるため，Hansen らの方法は細かい区間分割を好み，ほぼ全ての区間で事例数が k もしくはそれに近い数になる．見方を変えると，Hansen らの方法は k の設定に敏感であると言える．

3 提案手法

先述したように，提案手法では Kontokanen らのヒストグラム密度推定手法を用いて最適な区間の分割点を求める．Kontokanen らの方法では正規最大尤度（normalized maximum likelihood, 以下 NML）を最大化することを目標とし，NML から導かれるペナルティ項により，データへの適合度と簡潔さのバランスを取る．一方，計算時間短縮のため，本研究では NML の代わりにその近似である BIC [9] とよりペナルティ項の小さい AIC [1] を採用した [5]．各々の場合の提案手法を HistAIC, HistBIC と呼ぶ．提案手法では，最小の匿名度 k の制約を満たしながら，式 4, 5 の再帰スコア ($1 \leq e \leq m$) の下で動的計画法に従って，AIC もしくは BIC を最小化する．I は区間数 ($1 \leq I \leq m$)，ϵ は精度を表す．その他の記号は 2 節で導入したものと同じ意味である．

$$B_{I,e} = \min_{e'}\{B_{I-1,e'} - (n_e - n_{e'})\log(\epsilon \cdot (n_e - n_{e'})) - \log((c_e - c_{e'}) \cdot n)) + \text{Penalty}\} \quad (4)$$

$$B_{1,e} = -n_e \cdot (\log(\epsilon \cdot n_e) - \log((c_e - (x_1 - \epsilon/2)) \cdot n)) \quad (5)$$

[5] 予備実験では，NML を用いた場合と BIC を用いた場合では得られる区間分割の結果に大きな違いは見られなかった．

式 4 において e' は $I-1$ から $e-1$ まで動く．ペナルティ項 Penalty は BIC のとき $\frac{1}{2}\log n$，AIC のとき 1 となる．

例として，2 節で用いたものと同じ 1,000 個のデータに対して HistAIC, HistBIC を適用した分割結果を図 1 中段，下段にそれぞれに示す．左が最小の匿名度 k を 5 に設定した場合，右が k を 50 に設定した場合である．Hansen らの方法の結果と比較すると，提案手法は k の設定によらず，確率密度の形状を保った上でなるべく広い区間で分割することが見て取れる．特に，よりペナルティ項の大きい BIC を用いた方が区間幅は広くなる．

4 有用性の定量的評価

匿名加工後のデータの有用性を定量的に評価するために，本研究では密度比推定に基づく KL 情報量の推定値を用いる．具体的には，まず，匿名加工前のデータ（疑似識別子の集合）D と加工後のデータ D' の間で統計的な観点から変化した度合が少ないものほど匿名加工後のデータ D' は有用であると考える．そしてその度合を測るために，D と D' の各々が従う確率密度 p, p' 間の KL 情報量

$$\mathrm{KL}(p \,||\, p') = \int p(x) \log \frac{p(x)}{p'(x)} dx \quad (6)$$

を用いる．真の p と p' は不明であるため，例えば D と D' から各々の推定値 $\hat{p}, \hat{p}'$ を得て，$\mathrm{KL}(\hat{p} \,||\, \hat{p}')$ を近似的に用いることが考えられる．更に近年，p と p' を別々に推定せずに密度比関数 $r = p/p'$ を直接 $\hat{r}$ と推定し，

$$\widehat{\mathrm{KL}}(D \,||\, D') = \frac{1}{n}\sum_{i=1}^{m} w(x_i)\hat{r}(x_i) \quad (7)$$

として $\mathrm{KL}(p \,||\, p')$ を近似する方法が提案されている [13]．本研究ではその一つである uLSIF 法 [8] を実装した densratio[6] を使って $\widehat{\mathrm{KL}}(D \,||\, D')$ を求める．以降では $\widehat{\mathrm{KL}}(D \,||\, D')$ を「推定 KL 情報量」と呼ぶ．そして，推定 KL 情報量を匿名加工による情報損失量と見なし，これが小さいほど匿名加工後のデータは有用であると考える．

5 実験

5.1 実験の設定および手順

本研究では匿名度の平均が大きい区間分割ほど匿名性が高く，情報損失量が小さい区間分割ほど有用性が高いと考える．そして，Hansen らの方法と提案手法 (HistBIC, HistAIC) について匿名性と有用性のバランスを評価する．

評価用のデータとして，まず下の 3 種類の混合正規分布で生成した人工データを用意した．サイズは 1,000 と 10,000 の 2 通りである．

- $\mathcal{N}(1.0, 0.2^2)$, $\mathcal{N}(3.0, 0.5^2)$ を混合比 0.4, 0.6 で混合
- $\mathcal{N}(2.0, 0.6^2)$, $\mathcal{N}(9.0, 0.5^2)$ を混合比 0.2, 0.8 で混合
- $\mathcal{N}(1.0, 0.6^2)$, $\mathcal{N}(3.0, 0.8^2)$, $\mathcal{N}(8.0, 0.5^2)$ を混合比 0.2, 0.5, 0.3 で混合

また実データとして，UCI 機械学習リポジトリで提供される，1994 年米国の国勢調査データ（`https://archive.ics.uci.edu/ml/datasets/Adult`）の中の年齢属性のデータ（サイズ 48,842）を用意した．

実験の手順は次の通りである．まず，用意した各データを Hansen らの方法と提案手法 (HistBIC, HistAIC) で各々区

[6] `https://github.com/hoxo-m/densratio_py` から入手した．そして，より安定した結果を得るため，uLSIF 法におけるガウスカーネルの中心を匿名加工前のデータ D の最大値・最小値に囲まれた範囲で一様分布からサンプリングして得るように修正した．

表 1: サイズ 1,000 の人工データに対する匿名度の平均

最小の匿名度	Hansen	HistAIC	HistBIC
5	5.5	17.2	100.0
25	25.6	47.6	111.0
50	52.6	90.9	125.0
75	76.9	100.0	167.0
100	100.0	125.0	167.0

表 2: サイズ 10,000 の人工データに対する匿名度の平均

最小の匿名度	Hansen	HistAIC	HistBIC
50	50.5	123.0	417.0
250	256.0	384.0	666.0
500	526.0	667.0	909.0
750	769.0	1000.0	1000.0
1000	1111.0	1250.0	1250.0

間分割する．そして，分割された区間に対して 1 節で述べたマイクロアグリゲーションとデータ合成を行う．最後に，これらの匿名加工後のデータから推定 KL 情報量を求める．

5.2 匿名性の比較

サイズ 1,000 の人工データ 3 種類における匿名度の平均を表 1 に，サイズ 10,000 の人工データ 3 種類の匿名度の平均を表 2 に示す．また，年齢データにおける匿名度の平均を表 3 に示す．サイズ 1,000 と 10,000 の人工データでは同じ傾向が見られており，HistAIC は Hansen らの方法の約 1.5 倍，HistBIC は Hansen らの方法の約 2 倍の匿名度となった．年齢データでは，HistAIC は Hansen らの方法の約 2 倍，HistBIC は Hansen らの方法の約 2.5 倍の匿名度となった．つまり HistBIC, HistAIC, Hansen らの方法の順で匿名性は高くなった．区間分割結果をヒストグラムで表した図 1 の区間幅を見ても，この順で匿名性が高くなることが分かる．

5.3 有用性の比較

5.3.1 マイクロアグリゲーションの場合

指定された最小の匿名度 k に対し，区間分割後にマイクロアグリゲーションを施したデータにおける推定 KL 情報量を図 2 に示す．マイクロアグリゲーションでは，Hansen らの方法が一番小さい値になる場合が多かった．特に，k の値が小さい時に有用である．マイクロアグリゲーションは各区間中の値をその区間内の値の平均で全て置き換える．そのため，加工後は元データと平均値は変わらないが，データの散らばりは小さくなる．Hansen らの方法では各区間の事例数を k に近くなるようにデータを細かく分割する性質がある．そのため，k の値が小さければ散らばりの変化の度合が小さくなり情報損失量は非常に小さくなる．HistAIC，HistBIC は k への依存の度合が低く，区間を広くとるため，k が小さくても特別に情報損失量が小さくなることはない．

逆に，k の値を大きく設定すると，Hansen らの方法における情報損失量の値は HistAIC，HistBIC に近づき，有用性における Hansen らの方法の優位性は無くなった．この場合は，5.2 節で見たように匿名性の高い結果を生む提案手法は有効であると言える．

5.3.2 データ合成の場合

区間分割後にデータ合成を行ったデータにおける推定 KL 情報量を図 3 に示す．人工データに対しては，Hansen らの方法，HistAIC，HistBIC は同程度の値になっており，最も情報損失量が少ない手法はデータと k の値によって様々に変わる．この場合も匿名性の高い結果を生む提案手法は有効である．更に，人工データに対するデータ合成の情報損失量はマイクロアグリゲーションの約 1 割となり，年齢データでは約 6 割となった．このことから，区間分割を行った後の匿名加工としてはマイクロアグリゲーションよりデータ合成の方が情報損失量が少ないと言える．

表 3: 年齢データの匿名度の平均

最小の匿名度	Hansen	HistAIC	HistBIC
5	687.9	1684.2	2325.8
750	1162.9	2713.4	3488.7
1500	2220.1	4070.2	4884.2
2500	3052.6	3757.1	5426.9

なお，年齢データでの結果を見ると，最小の匿名度 k の値が小さく設定されたときに Hansen らの方法の推定 KL 情報量が非常に小さくなっている点が目に付く．このデータでは年齢は整数値で記載されており，19 歳から 47 歳まで 1 歳ごとに約 1,000 事例から約 1,400 事例が存在する．一方，表 3 にて Hansen らの方法における匿名度が $k = 750$ で 1162.9 であることから分かるように，$k = 750$ まではほとんどの区間では同じ年齢の事例しか存在していなかった．そのため，この範囲で Hansen らの方法における推定 KL 情報量が小さいといっても，匿名加工の役割を果たしていないことになる．

6 おわりに

本論文では，Kontokanen らのヒストグラム密度推定手法に基づき，単一の数値的な疑似識別子を含むデータに対して，有用性を保つ範囲で，なるべく高い匿名性を確保するように区間分割する手法を提案した．また，密度比推定に基づき推定された KL 情報量を用いて匿名加工データの有用性の定量的評価を行った．評価結果から，データ合成において高い匿名性が必要な場合に提案手法は有効であることが分かった．また，区間分割後の匿名加工方法としてはマイクロアグリゲーションよりデータ合成の方が情報損失量が少ないことが分かった．今後の課題としては他の実データへの適用や複数の数値的な疑似識別子を持つデータへの拡張が挙げられる．特に後者について，MDAV 法 [2]，V-MDAV 法 [10]，Mondrian 法 [7] 等が知られているが，例えば空間分割法として知られる Mondrian 法との組み合わせが考えられる．

参考文献

[1] Akaike, H.: A new look at the statistical model identification. IEEE Trans. on Automatic Control 19(6), 716–723 (1974)

[2] Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11, 195–212 (2005)

[3] Hansen, S.L., Mukherjee, S.: A polynomial algorithm for optimal univariate microaggregation. IEEE Trans. on Knowledge and Data Engineering 15(4), 1043–1044 (2003)

[4] Kameya, Y., Hayashi, K.: Bottom-up cell suppression that preserves the missing-at-random condition. In: Proc. of TrustBus-16. pp. 65–78 (2016)

[5] Kifer, D., Gehrke, J.: Injecting utility into anonymized datasets. In: Proc. of SIGMOD-06. pp. 217–228 (2006)

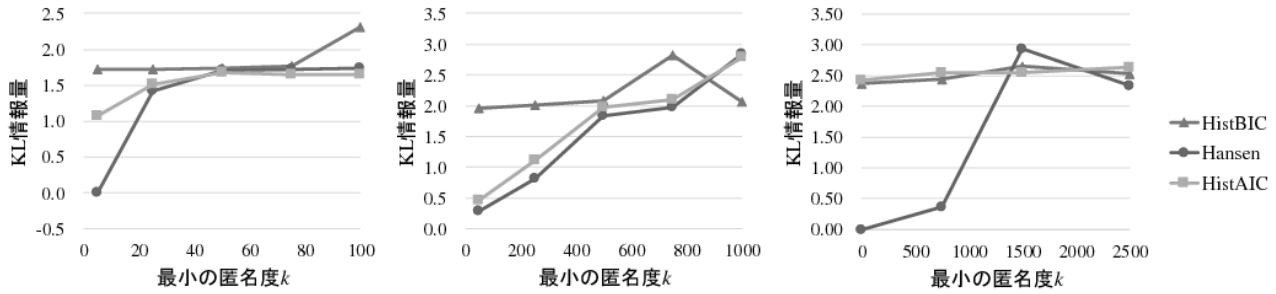

図 2: （左）サイズ 1,000 の人工データ,（中）サイズ 10,000 の人工データ,（右）年齢データに対してマイクロアグリゲーションを施したときの推定 KL 情報量.

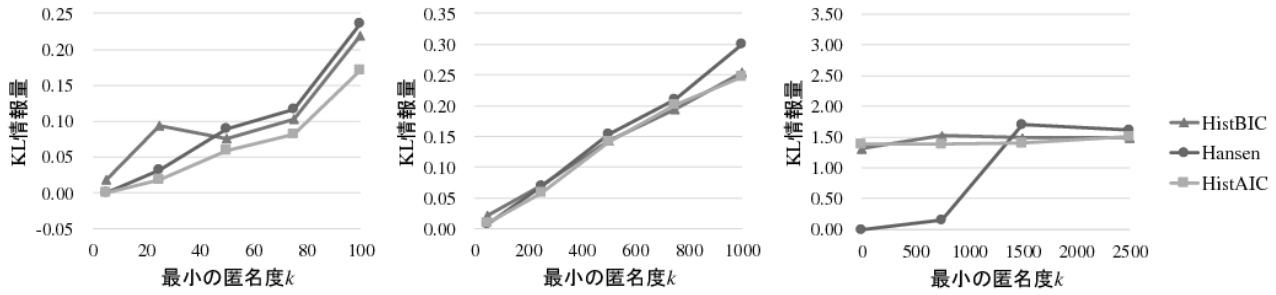

図 3: （左）サイズ 1,000 の人工データ,（中）サイズ 10,000 の人工データ,（右）年齢データに対してデータ合成を行ったときの推定 KL 情報量.

[6] Kontokanen, P., Myllymäki, P.: MDL histogram density estimation. In: Proc. of AISTATS-07. pp. 219–226 (2007)

[7] LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional *k*-anonymity. In: Proc. of ICDE-06. pp. 25–35 (2006)

[8] M. Sugiyama, T.S., Kanamori, T.: Density Ratio Estimation in Machine Learning. Cambridge University Press (2012)

[9] Schwartz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)

[10] Solanas, A., Martínez-Ballesté, A.: V-MDAV: a multivariate microaggregation with variable group size. In: Proc. of COMPSTAT Symposium of IASC (2006)

[11] Sweeney, L.: Achieving *k*-anonymity privacy protection using generalization and suppression. Int'l J. of Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 571–588 (2002)

[12] 秋山寛子, 和田昌昭, 中山雅哉, 加藤朗, 砂原秀樹: *k*-匿名化アルゴリズムにおける情報損失の極小化. 情報処理学会論文誌 57(12), 2675–2681 (2016)

[13] 杉山将: 確率分布間の距離推定：機械学習分野における最新動向. 日本応用数理学会論文誌 23(3), 439–452 (2013)

[14] 千田浩司: 個人特定のリスクを低減させる匿名化技術. オペレーションズリサーチ 61(5), 307–312 (2016)