6W-05

センシティブ属性値の距離を考慮したダミー追加による l-多様性アルゴリズムの提案

大石 慶一朗 † 清 雄一 ‡ 田原 康之 ‡ 大須賀 昭彦 ‡
電気通信大学情報理工学部総合情報学科 † 電気通信大学大学院情報理工学研究科情報学専攻 ‡

1 はじめに

個人に関する情報を含んだデータベースを他事業者と共有する場合，プライバシへの配慮が必要不可欠であり，既存指標の k-匿名性 [2] などに従って匿名化がなされてから共有されている. 既存指標の l-多様性 [3] は k-匿名性を拡張した指標であり，より良いプライバシ保護を実現できる. しかしながら，ある場合において l-多様性を満たしていたとしても事実上の多様性を満たしているとはいえない場合が存在する.

本研究では多くの既存研究と同様に，各個人の属性値ではなくデータベースの統計的な情報を他事業者が欲している状況を想定する．そのような状況で，事実上の多様性を満たせていない場合が存在するという課題を解決するために距離を考慮したカテゴリ分けを利用した，事実上の多様性を満たすことができるような手法及び指標を提案する.

2 既存指標及びその問題点

2.1 既存指標

k-匿名性 [2] はデータベースに対する匿名化を施した際，同一の準識別子 (Qid) の組み合わせが k 個以上存在することを保証するプライバシ保護指標である．対象となるデータベースは，個人を識別できる識別子，組み合わせることによって個人を識別することが可能な準識別子 (Qid)，個人の知られたくないプライバシ情報であるセンシティブ属性，その他の属性からなるものである.

k-匿名性を拡張した指標が l-多様性 [3] である．匿名化によって同一の Qid の組み合わせでグループ分けされた際，各グループ内におけるセンシティブ属性値が l 個以上であることを保証する指標である．表 2 は Qid の正確性を下げる匿名化の基本手法の一般化により二つのグループに分けられているが，どちらのグループも二種以上のセンシティブ属性値を保持しているので 2-多様性が満たされたデータベースといえる.

2.2 l－多様性の課題

一般化手法によって実現している l-多様性にはある問題が存在する場合がある．表 2 において仮 ID"C"と仮 ID"D"のグループに着目した場合，保持しているセンシティブ属性値は"胃癌"と"肺癌"であり，このデータベースは 2-多様性を満たしている. しかしながら，保持しているセンシティブ属性値がどちらも"癌"であるのでこのグループの人物は"癌"であることが確定してしまう. 仮 ID"C"が自身の病状が"癌"であることを知られたくない場合，このデータベースは事実上の多様性を満たしているとはいえなくなってしまう.

3 提案手法

3.1 提案指標

上で述べた課題を解決するために新しい指標，(l,d)– 意味的多様性を定義する.

(定義:(l,d)– 意味的多様性) ダミー追加後の匿名化データベースを T' とする．自然数 l 及び d に対して，T' が以下を満たすとき，T' は (l,d)– 意味的多様性を満たすという.

任意の自然数 i に対し，T' 内のレコード r_i が l 個のセンシティブ属性値を保持し，r_i のセンシティブ属性値間の最小距離 d_i が $d \leq d_i$ となる.

この指標を満たすことで距離の近いカテゴリに含まれる，類似したセンシティブ属性値の組み合わせができることを防ぐことが可能となる.

表 1 個人情報を含むデータベース

名前	性別	年齢	郵便番号	病状
桃子	女性	24	13000	盲腸
夏実	女性	22	13008	虫歯
五郎	男性	34	17025	胃癌
啓太	男性	34	17330	肺癌

表 2 一般化手法を用いた 2-多様性を満たすデータベース

仮 ID	性別	年齢	郵便番号	病状
A	女性	2*	1300*	盲腸
B	女性	2*	1300*	虫歯
C	男性	34	17***	胃癌
D	男性	34	17***	肺癌

Proposal of l-diversity algorithm considering distance between sensitive attribute values

keiichiro OISHI †, Yuichi SEI ‡, Yasuyuki TAHARA ‡, Akihiko OHSUGA ‡
† Department of Informatics, Faculty of Informatics and Engineering, The University of Electro-Communications
182-8585, Tokyo, Japan
‡ Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications
182-8585, Tokyo, Japan

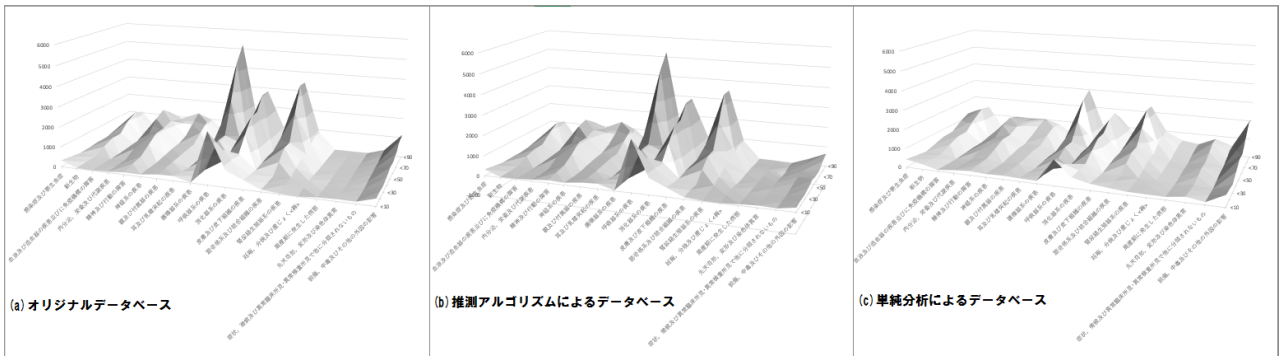

図 1　オリジナルデータセット及び提案アルゴリズムと単純分析で推測されたデータセットのヒストグラム

表 3　ランダム化手法を用いた 2-多様性を満たすデータベース

仮 ID	性別	年齢	郵便番号	病状
A	女性	24	13000	盲腸，肺癌
B	女性	22	13008	虫歯，盲腸
C	男性	34	17025	胃癌，虫歯
D	男性	34	17330	肺癌，胃癌

3.2　匿名化アルゴリズム

上記で定義した (l,d)– 意味的多様性に従って匿名化を行う. 本研究ではダミー追加手法を利用する. ダミー追加は Sei ら [1] の提案した手法で，Qid を一般化せずセンシティブ属性値にダミーを追加することで l-多様性を満たす手法である．表 3 はダミー追加によって 2-多様性を満たしている例であり，Qid より個人が識別されてしまったとしても 2-多様性を満たしているのでプライバシは保護されている．この手法を用いることでダミーが追加されるため各レコードの正確さは減少するものの，ダミー追加に対応した推測アルゴリズムを開発することにより，データベース全体から得られる統計的な情報に関しては，一般化手法よりも正確性が高いことが示されている．

本研究では一般化によるデータベースの有用性の低下も考慮しているのでこのダミー追加手法を利用する. ダミーを追加する際に保持してあるセンシティブ属性値を参照し，提案指標を満たせるカテゴリからダミーを選択し追加する.

3.3　推測アルゴリズム

提案手法ではカテゴリ分けを考慮してダミー追加を行っているので，提案手法に対する独自の分析手法が必要となる. 対象となるデータベース内の各センシティブ属性値を $(a_1,\cdots,a_F)$ とし，匿名化データベース内のセンシティブ属性値 a_i の総数を ω_i とする. その内のダミーではない真の a_i の総数を x_i，とすると以下の式が成り立つ.

$$\omega_i = x_i + \sum_{k\neq i} q_{(i,j)} \times x_k \tag{1}$$

また,$q_{(i,j)}$ はセンシティブ属性値 a_i が真値の時に a_j をダミーとして選択する確率であり，a_i を含むカテゴリ内のセンシティブ属性値総数を F_i とした際，次の式で求めることができる.

$$q_{(i,j)} = \frac{l-1}{F-F_i} \tag{2}$$

ただし，$i=j$ のときは $q_{(i,j)}=0$ である．式 (1) を利用することでセンシティブ属性値の真の数を推測することができる.

4　評価

4.1　評価に用いるデータ

使用したデータは Qid として年齢と性別，センシティブ属性として病状を保持している. レコード数 10 万，センシティブ属性値 8277 種類に対して $(2,3)$– 意味的多様性を行った. また，センシティブ属性値のカテゴリ分類は厚生労働省より公開されている ICD-10 に準拠し，階層構造になっているのを利用して距離を定義した.

4.2　評価指標

図 1 は年齢と病状のヒストグラムを表している．単純分析よりも提案した推測アルゴリズムの方が，より元のデータに近い分布であることがわかる．また，提案した分析手法と $x_i = \frac{\omega_i}{l}$ より求められる単純分析の平均二乗誤差 (MSE) をそれぞれ求めた結果，推測アルゴリズムによる MSE はは 0.0013，単純分析による MSE は 0.0381 となった．このことからも提案アルゴリズムがよい精度であることがわかる.

5　おわりに

本稿では，l-多様性の課題を解決する指標及び，センシティブ属性値を考慮したダミー追加を用いて l-多様性の課題を解決する手法の提案，実験を行った．今後の課題は厳密な距離の定義やより良い分析手法の提案，複数センシティブ属性への対応などを行っていく.

参考文献

[1] Sei, Y., Takenouchi, T., Ohsuga, A. *(l1, ..., lq)-diversity for Anonymizing Sensitive Quasi-Identiers*, Proc. IEEE Trustcom, pp.596-603, 2015.

[2] LeFevre, K., DeWitt, D. and Ramakrishnan, R., *Mondrian Multidimensional K-Anonymity*, Proc. IEEE ICDE, pp.25-25, 2006.

[3] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M., *L-diversity: Privacy beyond K-Anonymity*, ACM TKDD, Vol.1, No.1, pp.3-es, 2007.