5N-6

Pk-匿名化アルゴリズムの一改良法の検討

柿澤 美穂 * 渡辺 知恵美 † 古川 諒 ‡ 高橋 翼 §

概要

ランダム化を用いた匿名化手法の一つに Pk-匿名化 [1] [2] [3] がある．既存手法である Pk-匿名化は, k-匿名性を確率的指標に拡張した Pk-匿名性を保証するために, レコード所持者を 1/k 以上の確信度に絞り込めないように属性値にノイズを付与する．既存手法は, 元の属性値にラプラス分布に従ったノイズを付与することで Pk-匿名化を実現し, 所望の k の下で Pk-匿名性を満たすようにラプラス分布の分散値を決定している．本稿では, より小さい分散値で Pk-匿名化を実現するよう改良したアルゴリズムを提案する．提案手法を既存手法と比較, 評価することで提案手法の優位性を示す．

1 はじめに

近年, データベースサービスの普及に伴い, 個人情報等の機密情報をデータベースに格納する場合のプライバシ保護が要求されている．特に, データベースに格納された機密情報を公開する際, データ公開者はデータベースのレコード所持者をデータ利用者に特定させずに公開したいと望む．そのような場合にレコード所持者を隠すため, データを匿名化する手法が研究されている．データ匿名化の一手法として, k- 匿名化 [5] [4] という手法がある．k- 匿名化とは, 属性値の抽象化, 削除等を行うことによって, レコード所持者を k 人未満に絞れないようにする手法である．この k- 匿名化を確率的指標に拡張した手法として, Pk-匿名化 [1] [2] [3] がある．Pk-匿名化は属性値の置換やノイズ付与といった確率的な操作を用いて, レコード所持者を 1/k 以上の確信度で絞り込めないようにする．既存研究では数値属性に対して, ラプラス分布に従ったノイズをそれぞれ付与することで Pk-匿名化を実現する方法が提案されている．

既存手法の Pk-匿名化では, ラプラス分布の分散値を属性値間の最大距離に従って決定している．これにより, 全ての属性値を匿名化の範囲に含めることができる．しかし, 属性値の分布によっては, この方法では不必要なノイズが付与され過剰な匿名化がなされる場合がある．例えば, 幾つかのグループに分かれて属性値が存在しているような場合である．我々は, このような場合に全ての属性値間で最大距離をとるのではなく, 属性値のグループ毎にその中での属性値間の最大距離をとり, その最大距離に従いグループ毎に異なるパラメータでラプラス分布のノイズを付与すれば, 不必要なノイズ付与を防ぐことができると考える．そこで本稿では, これを既存手法の Pk-匿名化の一改良法として提案し, 既存手法との比較, 検証を行う．

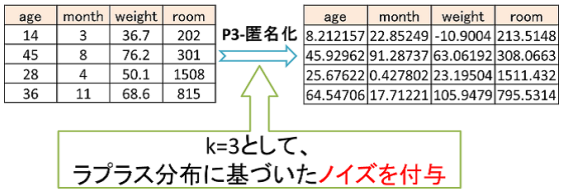

図 1 Pk-匿名化

2 Pk-匿名化とは

既存の Pk-匿名化は, レコードの数値属性に対し, ラプラス分布に従ったノイズを付与することで実現する．ここでは, k の値を下記の式で定義している．

$$k = 1 + (|R| - 1) \prod_{V:\text{属性}} exp(-2\frac{sup_{u,v\in V}||u-v||_1}{\sigma})$$

σ はラプラス分布の分散値である．所望の k の値の下で, σ の値を上記の式で決定する．

複数属性の場合は, まず各属性の値域を, 属性毎に [0,1] の範囲に正規化する．これは, ラプラス分布のパラメータの次元を属性間で揃え, 属性間でノイズの強度が異なるのを防ぐためである．正規化した後, 上記の式で σ の値を算出し, σ の値に基づいたノイズを付与してから正規化の逆変換で値のスケールを元に戻す．

3 既存手法の課題と提案手法

この時, 既存の k の算出式を変形すると, σ の値は属性値間の最大距離 $sup_{u,v\in V}||u-v||_1$ に依存していることが分かる．

* お茶の水女子大学 人間文化創成科学研究科
† 筑波大学 システム情報工学研究科
‡ 日本電気株式会社 クラウドシステム研究所
§ 日本電気株式会社 クラウドシステム研究所

$$\sigma = 2\frac{sup_{u,v\in V}||u-v||_1}{log(|R|-1)-log(k-1)}$$

元の値がある程度均等に分布している場合 (図 2), 属性値間の最大距離をとっても不必要なノイズが付与されることはない．しかし, 例えば元の値が二つに分類された分布を持つ場合 (図 3), 属性値間の最大距離をとると, 値が分布していない空間を含んでいるにも関わらずその最大距離で σ の値を決定するため, σ の値は必要以上に大きくなってしまう．

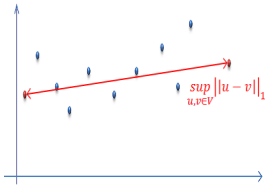

図 2　値が均等に分布している場合

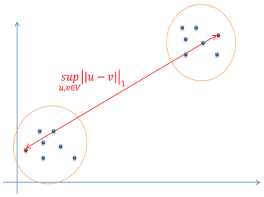

図 3　値が偏って分布している場合

我々の提案手法は, 属性値が偏って分布している場合に, ある一つの属性値とその周辺に分布している属性値をまとめて一つのグループとし, そのグループ毎に σ の値を算出しノイズを付与するという手法である．これにより, 必要以上のノイズが付与されることを防ぐ．

4　実験と評価

元の値をあらかじめ分類して Pk-匿名化を施すことにより, より小さい σ の値でのノイズ付与を実現することが期待できる．ここで, サンプルデータを用いて実験を行う．東京と大阪の全国地方公共団体コードと郵便番号が格納されたテーブルを Web 上からダウンロードし, 用意する．東京と大阪のレコードの分布は図 4 のようにはっきりと分かれており, 我々はこれを東京グループと大阪グループにあらかじめ分類したデータとして利用する．

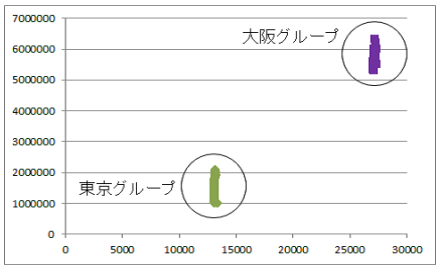

図 4　東京グループと大阪グループの分布

東京と大阪をまとめたデータを東京_大阪グループと呼ぶ．この東京_大阪グループに対して σ の値を求めた場合と, 東京グループ, 大阪グループで別々に σ の値を求めた場合との σ の値を比較する．図 5 が, σ の値を比較したグラフである．東京グループもしくは大阪グループだけで σ の値を求めた場合に比べて, 東京_大阪グループで σ の値を求めた場合の σ の値の方が 6〜7 倍大きくなっていることが分かる．

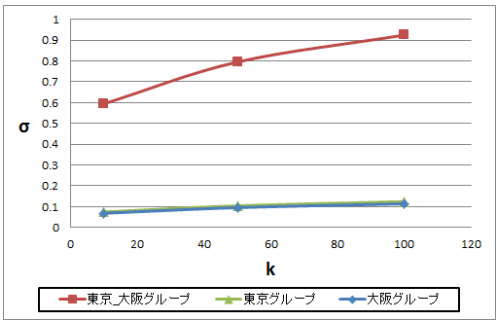

図 5　σ の値の変化グラフ

この結果より, 値の分布に偏りがある場合, あらかじめ分類をしてからグループ毎に σ の値を決定し, その σ の値に基づいてノイズを付与する方法が有用であることが言える．

5　まとめ

本稿では, ラプラス分布のパラメータを決める際に元の属性値の分布を考慮し, 元の属性値をあらかじめいくつかのグループへ分類してからグループ毎に Pk-匿名化を施すという, 一改良法を示した. この手法により, 過剰にノイズを付与することなく Pk-匿名化を実現することが可能になると考える．今後の課題として, さらに複雑な分類を持つデータや実際の機密情報を用いた検証実験や, k の値と σ の値との関係の数式的定義の確立などが考えられる．

参考文献

[1] 五十嵐 大, 千田 浩司, 高橋 克巳,”k-匿名化の確率的指標への拡張とその適用例”,CSS2009,2009

[2] 五十嵐 大, 千田 浩司, 高橋 克巳,”数値属性における k 匿名化を満たすランダム化手法”,CSS2011,2011

[3] 五十嵐 大, 千田 浩司, 高橋 克巳,”ランダム化データベース上の k-匿名性の一般的算出法”,CSS2011,2011

[4] 千田 浩司, 木村 映善, 五十嵐 大, 濱田 浩気, 菊池 亮, 石原 謙,”集合匿名化データの多変量解析評価”,CSS2012,2012

[5] L. Sweeney. k-anonymity: a model for protecting privacy.International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,10(5), pp.555-570, 2002.