データベースのプライバシを保護するための Pk-匿名化手法の精度改良に関する諸検討

理学専攻・情報科学コース　柿澤 美穂

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 [2] がある. k-匿名化とは, 属性値の抽象化や削除を行い, 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が, Pk-匿名化 [1] である. Pk-匿名化は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. しかし, 既存の Pk-匿名化手法には, ノイズが過剰に付与されているという課題がある.

そこで本研究では, 既存の Pk-匿名化手法の一改良法として, 属性値を予め複数のグループに分類してから Pk-匿名化を施す手法を提案する. 既存手法に比べて, グループに分類することでノイズが従うラプラス分布の分散を小さく抑えることができ, 結果として過剰なノイズを抑えることができる. さらに, 属性値のグループ分類に Mondrian と DBSCAN というクラスタリング手法を併用した方法を適用し, より効果的にノイズを抑えられることを示す.

2　Pk-匿名化

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 と σ の関係式を以下の様に定義している.

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

σ はラプラス分布の分散を示しており, ラプラス分布の広がり方を示す数値として扱う. 所望の k の値の下で, σ の値を上記の式で決定する.

Pk-匿名化によってどのようにプライバシが保護されるのかを, 例を用いて示す. ここに, 図 1 のような, ある病院の患者の氏名, 年齢, 体重, 病名が格納されているテーブルがあるとする. このテーブルを, 病院のスタッフがデータアナリストに公開し, アナリストは患者のプロフィールと病気の関係を分析するとする. ここで, 患者のプライバシを保護するため, データを公開する病院スタッフは, 匿名化処理として Pk-匿名性を満たすようなノイズを, レコードの各数値属性に付与する. そこにある攻撃者がいて, Alice の年齢が 13 , 体重が 36.7 という数値と, 匿名化後のテーブルを知ってしまっているとする. この攻撃者は, 自身の知っている値から Alice のレコードを判別し, Alice の病気を特定したい. この時,

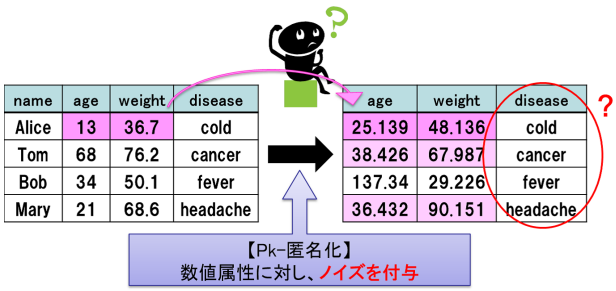

図 1: Pk-匿名化の例

攻撃者が匿名化前のテーブルを見ると, Alice の病気が分かってしまうが, テーブルが匿名化されていると, 値が変わっているために Alice に対応するレコードがどれなのか分からず, Alice の病気を特定することができない.

このようにして, レコードの値にノイズを付与し攪乱することで, プライバシを保護するのが Pk-匿名化である.

3　既存手法の課題

既存手法の課題として, Pk-匿名化後のデータの有用性は k-匿名化後のデータより低いことを指摘する. ここでは, k-匿名化データと Pk-匿名化データの相関係数を比較する実験を行い, 匿名化後の値が元の値の特徴をどれだけ保持して匿名化できているかを検証する. 対象データとして, レコード数 30000 , 相関係数が 0.5 であるランダムデータを使用し, 結果を以下のグラフ (図 2) に示す. k-匿名化データの相関係数は元データの相関係数と大きく異ならないのに対し, Pk-匿名化データの相関係数は元データの相関係数から大きく離れている. このことから, Pk-匿名化データは k-匿名化データよりも匿名化による値の分布の広がり方が大きく, k-匿名化データに比べてデータ有用性が低いということが言える.

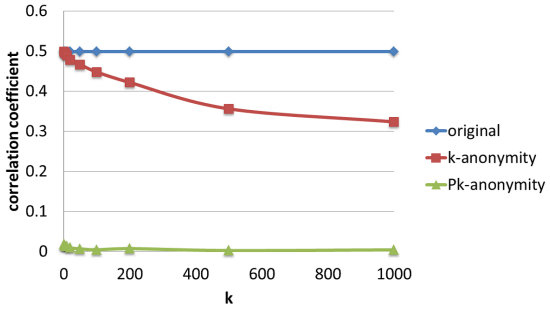

図 2: 相関係数 0.5 のデータの相関係数の推移

4　提案手法と実験

そこで本研究では, 既存の Pk-匿名化手法の改良法として, 元の属性値を予め複数のグループに分類し, グループ毎に Pk-匿名化を施す手法を提案する. 前述の式 (1) の分子は属性値間の最大距離を表しており, 属性値に付

与されるノイズのラプラス分布の分散は, 属性値間の最大距離に依存していることが分かる. そのため, 属性値をグループに分類することで最大距離が小さくなり, ラプラス分布の分散も小さく抑えられ, その結果過剰なノイズ付与を防ぐことができる.

属性値の分類方法として, 濃度ベースクラスタリングの一手法である Mondrian と DBSCAN を適用する. Mondrian は, レコードの中央値を決めて左右に分割する作業を再帰的に繰り返すため, 近くに分布する属性値を同じグループに含めることができる. DBSCAN は, ある基点から同じクラスタに含めることができる点を推移的にたどっていきクラスタを形成する. DBSCAN により, クラスタから大きく離れて分布する属性値を外れ値として除外することができるため, DBSCAN を適用した状態で Mondrian を併用すると, さらにラプラス分布の分散を小さく抑えられる. 実験として, ある 1000 個の属性値を持つサンプルデータセットに DSCAN と Mondrian を適用し, 属性値をグループ分類した.

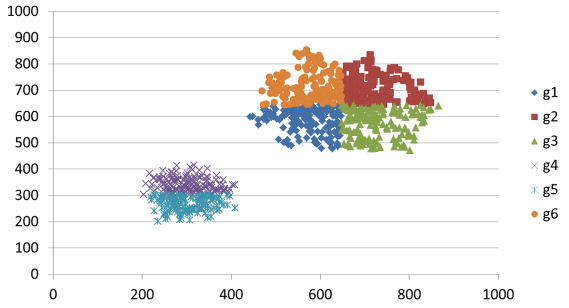

図 3: 提案手法:属性値を複数のグループに分類

図 3 のように 6 個のグループに分類され, 各グループ 140〜160 個の属性値を持つ. グループ毎に $k = 5$ として Pk-匿名化した場合の分散と, 既存手法を適用した場合の分散を全グループで比較した結果が, 以下の図 4 である.

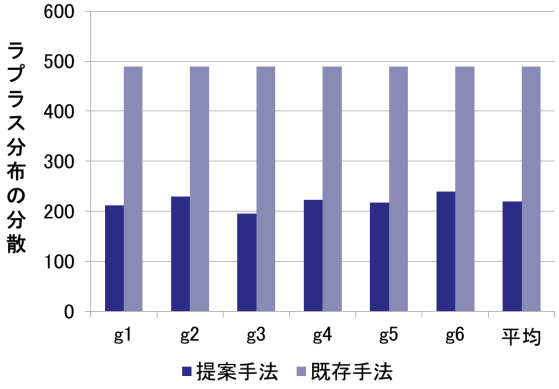

図 4: 既存手法と提案手法の分散の比較

この結果を見ると, グループに分類してから匿名化することで, 平均して約半分の分散で Pk-匿名化を実現できていることが分かる. 従って, 元の属性値を予めグループに分類してから, グループ毎に Pk-匿名化を施す方法が有用であると言える.

5 再構築法を用いたベイズ推定

さらに本研究では, Pk-匿名化により攪乱されプライバシ保護されたデータが, 統計的に有意なクロス集計結果を得ることができるかを検証するため, 再構築法を用いたベイズ推定を行い, クロス集計結果の比較を行う.

手順は, まず Pk-匿名化データからクロス集計を取得し, 反復ベイズ法を用いた再構築法によって付与されたノイズを薄めていく. クロス集計ベクトルの L1-距離が収束するまで繰り返す. Pk-匿名化において属性値がクロス集計のブロック間を遷移する確率は, 遷移確率行列によって表される. Pk-匿名化ではラプラス分布に従ったノイズを付与していることから, 遷移確率行列の成分を以下の様に定義し用いる.

$$A_{a,a} = 1 + \frac{\sigma}{w}(exp(\frac{-w}{\sigma}) - 1)$$

$$A_{a,a\pm i} = \frac{\sigma}{2w} exp(\frac{-w}{\sigma}(i-1))(exp(\frac{-w}{\sigma}) - 1)^2$$

ただし, $A_{a,a}$ は同じブロックに遷移する確率, $A_{a,a\pm i}$ は i 個隣のブロックに遷移する確率, w はクロス集計のブロック幅, σ はラプラス分布の分散を表す.

簡単のために, 定義域 10〜90 の 200 個のレコードを持つデータセットを用意し, $k = 2$ として Pk-匿名化しベイズ推定を適用する. クロス集計幅は 20 とし, 各レコードは値域に当てはまる 4 つのブロックに分類される. このデータセットにおいて, Pk-匿名化データと元データのクロス集計の L1-距離は 140 であるのに対し, ベイズ推定適用後のクロス集計と元データのクロス集計の L1-距離は 37.9 となった. 従って, ベイズ推定を適用することで, プライバシを保護しつつ元データのクロス集計に近い結果を出すことができ, Pk-匿名化データよりも統計的に有意な結果を得ることができる.

6 まとめと今後の課題

本研究では, Pk-匿名化手法における過剰なノイズ付与を防ぐため, 元の属性値を予め複数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提案した. 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. また, Pk-匿名化後のデータに対して再構築法を用いたベイズ推定を行うことにより, 匿名化後のデータからも統計的に有用なデータを取得することが可能であることを示した. 今後の課題として, 実際の大規模機密データへの提案手法の適用実験, 属性値の密度を考慮したグループ分類方法の提案等が挙げられる.

謝辞

本研究に対して熱心にご指導賜りました筑波大学の渡辺知恵美助教, 日本電気株式会社の森拓也氏, 古川諒氏, 高橋翼氏に心から感謝申し上げます.

参考文献

[1] 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

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