平成26年度 情報工学コース卒業研究報告要旨

関 研究室	氏　　名	中 島　聖 斗
卒業研究題目	プライバシ保護のための射影分解法の実装とその評価	

本研究では，プライバシ保護のモデルである t-近接性を実現するためのアルゴリズムである射影分解法の実装を行った．また，公開されている約3万人の成人レコードからなるデータ集合に対して実装システムを実行し，t-近接性を実現する別のアルゴリズムであるラプラス雑音法との比較を行うことで射影分解法の評価を行った．

t-近接性はより基本的なプライバシ保護基準である k-匿名性の拡張である．データ集合が従う属性集合が公開しても良い属性である準識別子と，準識別子と対応して公開されるべきではない属性である秘匿属性の集合に分割されていると仮定する．このとき，データ集合が k-匿名性を満たすとはどの準識別子の値についてもその値を持つレコードが k 個以上含まれることをいう．これにより特定の個人（ターゲット）の秘匿属性値を知りたい攻撃者は，ターゲットの秘匿属性値の候補を k 個より絞り込むことができない．データ集合が t-近接性を満たすとは，k-匿名性を満たしたうえで，準識別子の値が等しいデータ集合の各部分集合において，秘匿属性値の分布がデータ集合全体の秘匿属性値の分布と類似していることをいう．ここで2つのデータ集合の秘匿属性値の分布が類似しているとは，2つのデータ集合の距離の定義を与え，その距離が t 以下であることとする．

データ集合に対して t-近接性を満たすように変換を行うアルゴリズムである射影分解法では，バケッティングとパーティショニングの2つのステップで変換を行う．バケッティングでは秘匿属性値の種類を $b = t + 1$ 個以下に抑制する．パーティショニングでは，バケッティングを行ったデータ集合を b 個のグループに分割する．この時，各グループではある一つのバケットだけ個数が他のバケットの個数より t^2 倍含まれ，それ以外のバケットの個数は均等に含まれるようにする．

射影分解法の実装はRubyで行った．データ集合はCSV形式のテキストファイルで与えられ，t の値も入力として与える．はじめにデータ集合を秘匿属性値によって昇順にソートし，近い値が同じグループに含まれるように b 個のグループに分割する．そのグループをバケットとし各レコードの秘匿属性値をバケットの値で書換える．バケットの値として平均値を用いるが，比較のため，書換えを行わずそれぞれのレコードの秘匿属性値をそのまま用いる場合も考える．バケッティング後，i 番目のグループに i 番目のバケットのレコードが他のバケットのレコードより t^2 倍だけ多く含まれるように b 個のグループに分割する．得られたグループについて，元のデータ集合と射影分解後の各グループに対して秘匿属性値のヒストグラムを計算した．

また，t-近接性を満たすデータ集合は同時に，$t = \exp(\varepsilon)$ において ε-差分プライバシも満たすことが知られている．そこで，ε-差分プライバシを実現するアルゴリズムであるラプラス雑音法の実装を行い，比較実験を行った．射影分解法ではバケットの値として秘匿属性値をそのまま用い，得られるグループに対する問合せ結果がランダムに変化するように設定をし，射影分解法によるグループに対する問合せ結果の分散と射影分解法を一度実行して得られたグループ対するラプラス雑音法による問合せ結果の分散を比較することで，射影分解法の評価を行った．データ集合に対する問合せは平均値を用いた．

その結果，ラプラス雑音法ではプライバシ保護の尺度である ε の値を小さくするにつれて加えられるノイズが大きくなり問合せ結果が大きく分散するのに対し，射影分解法では ε の値によらず一定の分散となることが確認できた．