修士論文 要旨

集合値データに対する個人適応型匿名化手法

数理情報学専攻 48166223　中川 拓麻
指導教員　中川 裕志 教授

1　序論

ビッグデータの利活用が重視される中，特に幅広い活用が見込まれるデータ形式として，集合値データ（set-valued data）がある．これはデータに含まれる一つ一つのレコードが集合として表されるものであり，例えば時系列的な順序構造を除いた Web 閲覧履歴や，e-commerce サイトにおける各顧客の購買履歴などを記録する際に用いられるデータ形式の一つである．このようなデータを広くビジネスや研究の現場で活用するため，第三者にデータを渡したり，一般に公開したいという需要がある．しかし，これはデータ主体のプライベートな情報の漏洩に繋がる恐れがあり，安易に実行することはできない．この課題を解決するため，公開したいデータの有用性を維持しつつプライバシーを守るようなデータ加工を行うための手法が，プライバシー保護データ公開（Privacy-Preserving Data Publishing: PPDP）の文脈で広く研究されている．

集合値データの安全性を保証するための匿名化モデルはこれまでに複数提案されている．しかし，既存のモデルでは，守ることのできるアイテムは事前に機微アイテムと指定した一部のアイテムに限られてしまうという課題があった．そこで本研究では，ρ-不確実性をベースとしながら，個人ごとに異なるアイテムを機微情報と見なすことを許し，これを「機微制約」として明示的に与えることで，現実的な状況において適用可能な匿名化モデル，個人適応型 ρ-不確実性を提案する．

2　匿名化モデル

Cao ら [2] によって提案された既存のモデル，ρ-不確実性を元に，提案モデルを以下のように定義する．考える集合値データセットを D，データセットに現れる全アイテムの集合（ドメイン）を $\mathcal{I}$ とする．ユーザー集合を $\mathcal{U}$ とし，ユーザー $u \in \mathcal{U}$ のレコードを D_u と書く（$D_u \subseteq \mathcal{I}$）．$D$ においてアイテム集合 $Q \subseteq \mathcal{I}$ を含むレコード数を Q のサポート（support）と呼び，$supp_D(Q)$ と書く．互いに重なり合わない 2 つのアイテム集合 Q，R（$Q, R \subseteq \mathcal{I}$, $Q \cap R = \emptyset$）について，相関ルール $Q \to R$ を考え，その確信度（confidence）を

$$conf(Q \to R) = \frac{supp_D(Q \cup R)}{supp_D(Q)} \tag{1}$$

と定義する．各ユーザー u に対し，ユーザー u が機微情報と見なすアイテムの集合を $E_u \subseteq \mathcal{I}$ とし，それらをまとめた $E = \{(u, E_u) : u \in \mathcal{U}\}$ を機微制約と呼ぶ．攻撃者 $adv = (u, Q)$ はユーザー u（ターゲット）がアイテム集合 $Q \subseteq D_u$ を持っているということを知っており，この情報を用いて，$e \in E_u \backslash Q$ を u が持つ確率を $conf(Q \to \{e\})$ によって推定する．事前に定めた，プライバシー強度の基準を表すパラメータ ρ（≤ 1）を用いて，満たすべき条件を次のように定める．

定義 1. データセット D は，攻撃者 $adv = (u, Q)$ について次のいずれかが成り立つとき，adv に対して安全であるという：

1. $\forall e \in E_u \backslash Q$，$conf(Q \to \{e\}) \leq \rho$.
2. $supp_D(Q) = 0$.

この条件を用いて，データベース全体のプライバシーを次のように定義する．

定義 2. データセット D は，任意の攻撃者 $adv = (u, Q)$ に対して安全であるとき，個人適応型 ρ-不確実性を満たすという．

3　データ加工アルゴリズム

本研究では集合値データを加工する手法として，データ中のアイテムの抑圧（削除）を考える．抑圧の手法は，各アイテムを全レコードから属性ごと抑圧する大域的抑圧，各アイテムを一部のレコードにおいてのみ抑圧することを許す局所的抑圧に大別される．本研究ではこのそれぞれについて，定義 2 の条件を満たしつつ，以下によって定義される有用性損失の指標を最小化するためのアルゴリズムを構築した:

$$util_{\text{info}}(D, D') = \frac{\sum_{i \in \mathcal{I}} (supp_D(i) - supp_{D'}(i))}{\sum_{i \in \mathcal{I}} supp_D(i)}. \tag{2}$$

ここで，D は加工前のデータ，D' は加工後のデータを表す．

3.1 大域的抑圧

最適な大域的抑圧を行う問題は，線形な整数計画問題として定式化して解くことができる．また，既存研究における類似した問題に対して，従来から貪欲法を用いた効率的解法 [2, 4] が提案されていたが，本研究ではこの手法を提案モデルに適用可能な形に拡張した．さらにこれについて，等価な集合被覆問題へと帰着できることを示し，その近似率について理論保証を与えた．

3.2 局所的抑圧

最適な局所的抑圧を行う問題についても，Gkoulalas-Divanis ら [3] による線形化の手法を用いることで，線形な整数計画問題として解いて厳密解を得る手法を示した．また，プライバシー侵害のリスクを反復的に除いていくことによってデータ全体の安全性を保証するヒューリスティックアルゴリズムを構築した．

4 確率的緩和モデル

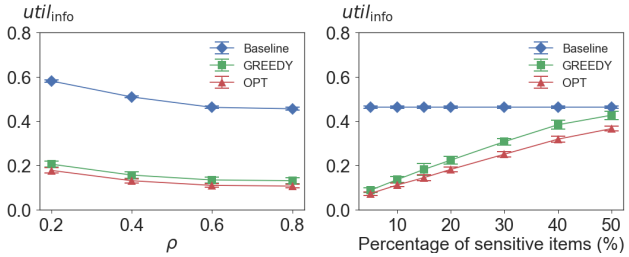
これまでのモデルは，安全性を保証するために必要な計算コストがデータの最大レコード長に対して組み合わせ的に増加するという本質的な課題を抱えていた．これに対し，Gergely ら [1] によるアイデアを用いてモデルを確率的に緩和することによって対処する．

いま，ターゲットの持つレコードの大きさ ℓ の部分集合を背景知識として持つ攻撃者の集合を $\mathcal{A}^\ell = \{(u, Q) : u \in \mathcal{U}, Q \subseteq D_u, |Q| = \ell\}$ と書き，α_ℓ を $\mathcal{A}^\ell$ 上に値をとる確率変数とする．また，D が α_ℓ に対して安全でない確率を H_ℓ とする．

定義 3. データセット D 及び機微制約 E を考える．すべての $\ell \leq m$ について，$Pr[H_\ell < \varepsilon] \geq 1 - \delta$ が成り立つとき，D は個人適応型 (ε, δ)-ρ^m-不確実性を満たすという．

十分小さい $\varepsilon, \delta > 0$ についてこの定義が満たされるとき，D は高い確率で大半の攻撃者に対して安全であるということが言える．次の定理に基づき，この定義を満たすための加工アルゴリズムを，大域的抑圧，局所的抑圧のそれぞれの場合について構築した．

定理 1. データセット D, 機微制約 E において，α_ℓ が従う確率分布から独立にサンプリングされたサンプル集合を S とする．$\varepsilon, \delta > 0$ について，$|S| \geq \frac{\log(1/\delta)}{2\varepsilon^2}$ が満たされているとする．このとき，すべての $adv = (u, Q) \in S$ について D が安全ならば，$Pr[H_\ell < \varepsilon] \geq 1 - \delta$ が成り立つ．



(a) 機微アイテムの割合 = 10%　　(b) $\rho = 0.5$

図 1: BMS-WebView-2 [5] における結果

5 実験

Web 閲覧履歴及び購買履歴の実データ [5] を用いて，提案アルゴリズムの性能を調べる実験を行った．図 1 には，個人が異なるアイテムを機微アイテムとして守りたいと指定している状況で，大域的抑圧によって定義 2 を満たすように加工を行った結果を示す．Baseline は既存の手法 [2]，OPT は最適な加工，GREEDY は貪欲アルゴリズムによる加工の結果を示している．Baseline においては，少なくとも一人以上のユーザーにとって機微であるアイテムはすべてのユーザーにとって機微であるものとして扱うこととになり，有用性損失が非常に大きくなる．一方で，提案手法を用いると大幅に結果が改善されていることがわかる．貪欲アルゴリズムを用いた場合も，最適値に近い結果が得られている．

参考文献

[1] Gergely Acs, Jagdish Prasad Achara, and Claude Castelluccia. Probabilistic km-anonymity efficient anonymization of large set-valued datasets. In *Proceedings of the 2015 IEEE International Conference on Big Data*, pages 1164–1173, 2015.

[2] Jianneng Cao, Panagiotis Karras, Chedy Raïssi, and Kian-Lee Tan. ρ-uncertainty: inference-proof transaction anonymization. In *Proceedings of the VLDB Endowment*, volume 3, pages 1033–1044, 2010.

[3] Aris Gkoulalas-Divanis and Vassilios S. Verykios. An integer programming approach for frequent itemset hiding. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 748–757, 2006.

[4] Yabu Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. Anonymizing transaction databases for publication. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–775, 2008.

[5] Zijian Zheng, Ron Kohavi, and Llew Mason. Real world performance of association rule algorithms. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 401–406, 2001.