ERATO セミナ 2010 - No. 05

データ匿名化に関する検討

白井康之
JST ERATO 湊離散構造処理系プロジェクト

2010/7/9

概要

個人情報をはじめとする機微なデータから個人が特定されないようにする技術は，近年の個人の行動履歴情報の利活用の進展に伴い，特にデータマイニングの分野で注目を集めている．k 匿名性をはじめとする匿名化技術は，基本的にはデータ項目の組み合わせ問題に帰着するが，匿名化されたデータの効率的な管理という側面から見れば，ZDD をはじめとしたデータベースの効率的な管理手段が必要とされる分野である．本発表では，データの匿名化が必要とされる背景や研究動向等のサーベイを行った上で，必要とされる匿名化技術の概要ならびに ZDD によるコーディング手法に関する検討結果について触れる．

データ匿名化に関する検討

白井 康之
ERATO Minato Discrete Structure Manipulation System Project
Japan Science and Technology Agency
shirai@erato.ist.hokudai.ac.jp

目　次

概　要

データの匿名化とは何か？

- 事業者はデータ（個人情報）を持っている．
- 事業者間のデータをマージすると，より意味のあるデータを作ることができる．
- しかし，個別の個人情報が流出するのは問題．

- 個人情報とは，氏名，住所，生年月日等，個人を「直接的に」特定できる情報
- しかし，実は直接的に特定できなくても「間接的」に識別できる情報が存在．
- たとえば，ある特定地域で奇病にかかっている人（ひとりしかいない！）など

- そこで...
  - データを流通させる場合には，情報の匿名化が必要である．
  - データを流通させずに，分析結果のみを開示できる仕組みが必要である．
- 前者は個人情報匿名化，後者はセキュア計算と呼ばれる．
- Privacy Preserving Data Mining という言葉は両者を含んでいる．

たとえば...

識別情報（個人を特定可能な情報）

氏名	電話番号	年齢	性別	趣味	疾患	年収
山田	03-...	42	男	車	糖尿病	500
鈴木	048-...	32	男	映画	糖尿病	600
小林	03-...	34	男	映画	糖尿病	620
田中	03-...	41	男	旅行	糖尿病	600
山本	03-...	42	男	旅行	なし	450
佐藤	03-...	28	女	旅行	糖尿病	700
井上	03-...	24	女	旅行	糖尿病	600

識別子を持たない情報
（だが，《年齢と性別から》個人が特定可能）

氏名	電話番号	年齢	性別	趣味	疾患	年収
山田	03-...	42	男	車	糖尿病	500
鈴木	048-...	32	男	映画	糖尿病	600
小林	03-...	34	男	映画	糖尿病	620
田中	03-...	41	男	旅行	糖尿病	600
山本	03-...	42	男	旅行	なし	450
佐藤	03-...	28	女	旅行	糖尿病	700
井上	03-...	24	女	旅行	糖尿病	600

準識別子

たとえば...

識別情報（個人を特定可能な情報）

氏名	電話番号	年齢	性別	趣味	職業	家族構成
山田	03-...	42	男	車	自営業	独身
鈴木	048-...	32	男	映画	会社員	既婚
小林	03-...	34	男	映画	会社員	既婚
田中	03-...	41	男	旅行	無職	独身
山本	03-...	42	男	旅行	会社役員	既婚
佐藤	03-...	28	女	旅行	無職	既婚
井上	03-...	24	女	旅行	会社員	独身

識別子を持たない情報
（個人は特定可能だが...）

氏名	電話番号	年齢	性別	趣味	職業	家族構成
山田	03-...	42	男	車	自営業	独身
鈴木	048-...	32	男	映画	会社員	既婚
小林	03-...	34	男	映画	会社員	既婚
田中	03-...	41	男	旅行	無職	独身
山本	03-...	42	男	旅行	会社役員	既婚
佐藤	03-...	28	女	旅行	無職	既婚
井上	03-...	24	女	旅行	会社員	独身

PPDM（広義）の研究

➢ 定義はかなりばらばら...

- プライバシー保護データ公開
  - ランダム化（マイニングに影響を与えないようなノイズを挿入）
  - 削除（マイニング結果に影響を与えないレアなデータを削除）
  - あいまい化（k-anonymity, l-diversity, …）
  - あいまい化の計算量はNP困難. 近似解法, 効率化が課題.
  - 撹乱データ上でのマイニング手法をどうやるか.
- データマイニング結果の加工
  - 相関ルールの圧縮、クラスタの結合など
- 検索問い合わせ結果の加工
  - クエリの結果の撹乱
- 分散したプライバシー情報の暗号化処理
  - セキュア計算（セキュアマルチパーティプロトコル）

➢ 狭義には、セキュア計算をPPDMと呼び、最初の3つは、privacy preserving data publishing, statistical disclosure control と呼ばれることもある。(むしろそのほうが多い感じ)

➢ 以下では, 主に, プライバシー保護データ公開について記述.

公開データ利用に関するニーズと現状

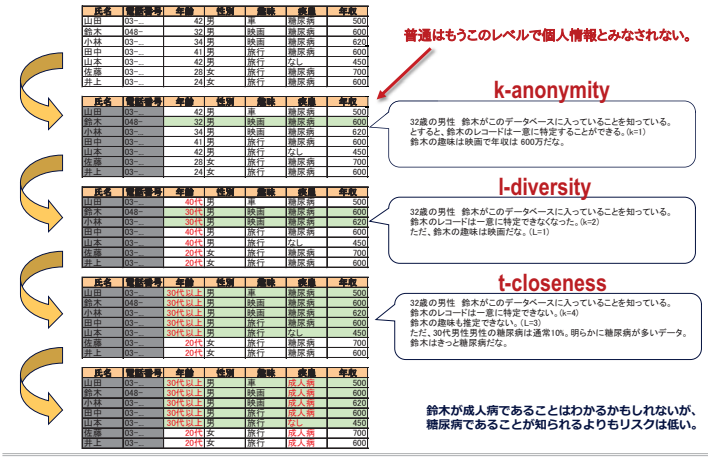
- 医療情報
  - 投薬情報, 病歴情報を共有し, 医療機関, 薬メーカ等で利用.
  - 特に薬剤メーカでニーズが高く, 一部実現している（手で対応）.
- クレジット情報
  - 過去のクレジット履歴を共有し, クレジットカード会社等で利用（個人を特定する形で既に実現）
  - 既存の枠組みを超えた情報共有も今後考えられる. たとえば, クレジット会社以外での利用など.
- 購入履歴情報･サービス利用履歴
  - 過去の他人の利用履歴から, リコメンデーションを提示.
    （これを買った人はこれも買っています）
  - 単一店舗でのリコメンデーションは実現されているが, 危ういケースも.
  - 一部データ販売サービスとして実現.

いくつかの研究分野（１）

- 収集データの公開手法に関する研究
  - データ公開の「基準」（何を満たせば公開して良いか?）
    - k-anonymity
      あるレコードに対して, 少なくとも他のk-1 レコードと区別できないことを保証する.
    - l-diversity
      属性のバリエーションがl 以上であることを保証する.
    - t-closeness
      グローバルな（既知の）分布との距離が t 以下であることを保証する.
    - l-diversity, t-closeness は現実問題としては細かすぎるとの議論もある.
  - 以下のような方法がある.
    - データを消す
    - データを変える（抽象化, スワッピング, アグリゲーション）
    - データを追加する（ランダムなデータの挿入）

参考：Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Springer, 2008

補足：匿名化の考え方



いくつかの研究分野（2）

- 利用価値を保存するためのデータ変換
  - 利用価値を定義する必要がある.
    「データの精度」ではほとんど意味がない.
  - privacy と accuracy にはトレードオフが存在する.
    （当たり前）
  - このトレードオフ関係は, 匿名化、抽象化の方法と, 適用するアルゴリズムによって決まってくる.
  - 目的は, privacy の基準を制約としてutility を最大化すること.
  - 解析手法により, さまざまな utility based PPDM が提案されている.

参考：Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Springer, 2008

いくつかの研究分野（3）

- プライバシー制約のもとでのマイニング手法の研究
  - 匿名化というよりはマイニングの研究.
  - 数ある手法に依存するPPDMだが, とてもホットな分野となっている（らしい）.
  - 2つの研究課題
    - 撹乱されたデータから, どのように正しい相関ルールが導出されたことを保証するのか.
    - 生成された相関ルールのプライバシー問題
      association rule hiding
      （例）
      「埼玉県在住で横浜市の運送会社に勤務する人は体重70kgである」
      （特定可能. もしくは高い確率で推定可能）

参考：Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Springer, 2008

いくつかの研究分野（4）

- 情報共有のための暗号化手法
  - いわゆる秘密計算，セキュア計算，セキュアマルチパーティプロトコルと呼ばれる分野．
  - データを直接触ることなく，マイニング結果のみを得る．
  - 簡単にいえば，暗号化された計算と計算結果の復号化．
  - Secure Sum, Secure Comparison, Secure Dot Product, Secure Union, …の組み合わせにより，相関ルールマイニング，決定木，k-NN分類器，ナイーブベイズ，SVM等を構成する手法が検討されている．
  - 究極の解といえるが，計算時間がかかるのが課題．
  - 国内では筑波大学の佐久間淳准教授など．

参考：Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Springer, 2008

いくつかの研究分野（5）

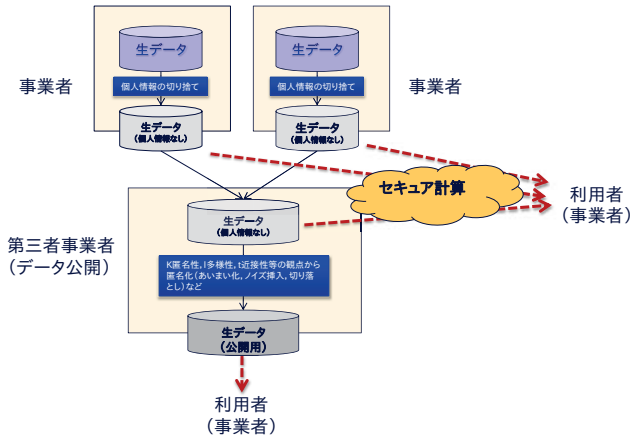
- 保護すべき情報の優先順位
  - さまざまな理由により，個人が高い確率で特定されないことを保証されなければならない個人と，必ずしもそうでもない個人が存在する．
    たとえば，風邪の患者と奇病の患者など．風邪の患者がばれてもたいした問題ではないが，奇病の患者については重大な問題。
- データストリーム上のPPDM
  - バッチ的に処理するのではなく，ストリームに対する匿名化処理．どこまで匿名化すればよいのかが判断できない．一度匿名化してしまうと復元できない，など．
  - Stream Data Mining と同じような難しさか．

参考：Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Springer, 2008

研究対象とする技術の候補

- Statistical Control for Disclosure Control
  k-匿名化によるデータの抽象化
- Utility Based Privacy Preserving DM
  privacy の基準を制約としてutility を最大化すること
- Personalized Privacy Preservation
  属性がばれても問題ないレベルとは？
- PPDM on Data Streams
  ストリームに対する匿名化処理

データマイニングとプライバシー保護（まとめ）



k-匿名化に関する既存研究（抜粋）

- P. Samarati, L. Sweeney, Generalizing data to provide anonymity when disclosing information, In Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 1998
- L. Sweeney, k-Anonymity: A Model for Protecting Privacy, Int'l Journal on Uncertainty, Fuzziness and Knowledgebased Systems, vol. 10, no. 5, 2002
- Kristen LeFevre, et al. (University of Wisconsin), Incognito: efficient full-domain K-anonymity, In Proc. of the 2005 ACM SIGMOD international conference on Management of data table of contents, 2005
- [広島市立大学] 村本，上土井，若林ほか，k匿名性を利用したデーター般化によるプライバシー保護，DEWS2007 など
- [情報大航海 2009] 日本情報処理開発協会，産総研情報セキュリティ研究センター，NTT情報流通プラットフォーム研究所，三菱総合研究所らは，共同で，情報大航海プロジェクトにおいて，k-匿名化に関する匿名化基盤ソフトウェアを構築し，実データを用いた検証を実施（2009年）．

匿名化の方法

準識別子（準識別情報）

準識別情報（個人を間接的に特定できる可能性がある）
《他のデータベースから参照可能》

識別情報（個人を直接的に特定できる）
《氏名，住所，電話番号など》

氏名	電話番号	年齢	性別	趣味	疾患	年収
山田	03-...	42	男	車	糖尿病	500
鈴木	048-	32	男	映画	糖尿病	600
小林	03-...	34	男	映画	糖尿病	620
田中	03-...	41	男	旅行	糖尿病	600
山本	03-...	42	男	旅行	なし	450
佐藤	03-...	28	女	旅行	糖尿病	700
井上	03-...	24	女	旅行	糖尿病	600

属性情報（個人を特定するキーにはならない）
《対象とするデータベースに固有の情報》

準識別情報の選択もひとつの課題

k-anonimity [Definition]

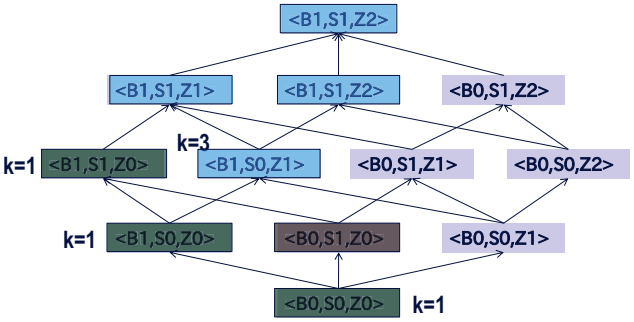
Quasi-Identifier《準識別子》
他のテーブルと結合させることにより，[高い確率で]個人が特定可能な属性の集合

k-anonymity Property
関係Tが属性集合Qに対して k-anonymity を満たすとは，同一のQに関する頻度が k またはそれ以上であること

k-anonymization《k-匿名化》
関係Tのk-anonymization とは，k-anonymity property を満たすように，T の Quasi-identifier に対して変更･削除を行うこと

K-anonymity (Incognitoの方法)

Kristen LeFevre, David J. DeWitt and Raghu Ramakrishnan, "Incognito: Efficient Full Domain K-Anonymity", In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data



・階層の一般化依存関係をグラフ化
・与えられたk 以上の部分については，no good として探索を省略

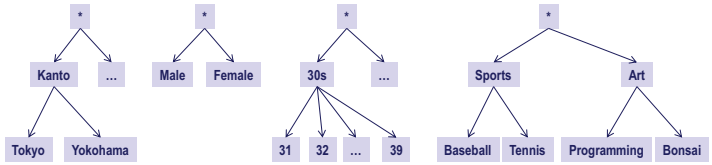
ZDDを使った表現方法の検討

ZDDを使った表現方法の検討

《ポイント》
- コンパクトな表現
  - BDD/ZDDを用いたコンパクトな表現
- メンテナンス性の向上
  - ➢ データの追加に対応
  - ➢ k匿名化への対応
  - ZDDベースのデータ管理手法
- 利用価値をベースにした最適な変換方法
  - VSOPを用いたコスト概念の導入

ZDDを使った表現（例）

ID	Place	Sex	Old	Hobby
A	Tokyo	Male	35	Baseball
B	Tokyo	Female	38	Bonsai
C	Yokohama	Male	36	Programming
D	Yokohama	Female	32	Tennis



place, sex, old, hobby すべてを準識別子とする。
準識別子以外のアイテムは各レコードに付随するデータとする(ので今回は省略)。

ZDDを使った表現（例：シンボルの定義）

◆変数の順序はここでは不問にする。
◆シンボル変数に値を挿入して、コストを表す。
◆「コスト」は秘密度合いを現す．
　ヒューリスティックなもの，または解析結果として定義されるもの．
　頻度とは若干意味が異なる．
◆最終的には，公開する・しないの判断をコストに委ねたい．

```
symbol   tokyo(1), yokohama(1), kanto(0), area(0)
symbol   male(2), female(2), gender(0)
symbol   y32(2), y35(2), y36(2), y38(2), y30s(1), years(0)
symbol   baseball(2), tennis(2), sports(1)
symbol   programming(3), bonsai(5), arts(3)
symbol   sports(3), arts(3), hobby(0)
```


ZDDを使った表現（例：データの定義）

```
S = 0
S = S + tokyo * male * y35 * baseball
S = S + tokyo * female * y38 * bonsai
S = S + yokohama * male * y36 * programming
S = S + yokohama * female * y32 * tennis

# Sに対して、年齢を不問にし、趣味をカテ
# ゴリ化 (必要条件として1つしかない項目
# を除外する)

T = 0
T = T + tokyo * male * y30s * sports
T = T + tokyo * female * y30s * arts
T = T + yokohama * male * y30s * arts
T = T + yokohama * female * y30s * sports
```

```
# Tに対して、趣味のみを残す
U = T
U = U / tokyo * kanto + U / yokohama * kanto
U = U / male * gender + U / female * gender

# Tに対して、性別のみを残す
V = T
V = V / tokyo * kanto + V / yokohama * kanto
V = V / sports * hobby + V / arts * hobby


# Tに対して、地域のみを残す
W = T
W = W / male * gender + W / female * gender
W = W / sports * hobby + W / arts * hobby
```


VSOPでの実行結果　（1）

目的：kの値が与えられた条件以上で，(maxcover, mincover) が大きい匿名化プラン

整数値ごとの列挙

```
/case S= 1:  tokyo male y35 baseball + tokyo female y38 bonsai + yokohama
male y36 programming + yokohama female y32 tennis
```

コスト最大の組み合わせ [コスト]
コスト最小の組み合わせ [コスト]

```
/maxcover S= (sensitive)
<Items>: tokyo female y38 bonsai  [10]
/mincover S= (unsensitive)
<Items>: yokohama female y32 tennis  [7]
```

```
/case T=1:  tokyo male y30s sports + tokyo female y30s arts + yokohama male
y30s arts + yokohama female y30s sports

/maxcover T= (sensitive)
<Items>: yokohama male y30s arts  [7]
/mincover T= (unsensitive)
<Items>: yokohama female y30s sports [5]
```


VSOPでの実行結果　（2）

住所と性別を一般化
k=2
maxcover=4
mincover=2

```
/case U= 2:  kanto gender y30s sports + kanto gender y30s arts

/maxcover U= (sensitive)
<Items>: kanto gender y30s arts [4]
/mincover U= (unsensitive) [2]
<Items>: kanto gender y30s sports
```

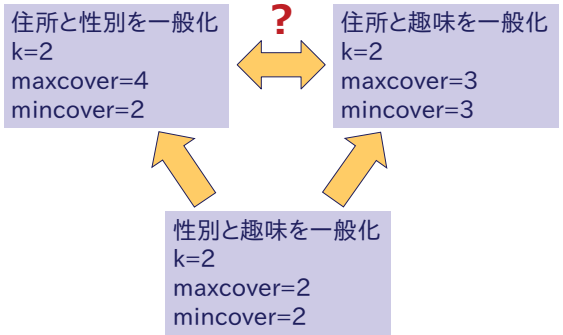
住所と趣味を一般化
k=2
maxcover=3
mincover=3

```
/case V= 2:  kanto male y30s hobby + kanto female y30s hobby

/maxcover V= (sensitive)
<Items>: kanto female y30s hobby  [3]
/mincover V= (unsensitive)
<Items>: kanto female y30s hobby [3]
```

性別と趣味を一般化
k=2
maxcover=2
mincover=2

```
/case W= 2:  tokyo gender y30s hobby + yokohama gender y30s hobby

/maxcover W= (sensitive)
<Items>: yokohama gender y30s hobby [2]
/mincover W= (unsensitive)
<Items>: yokohama gender y30s hobby  [2]
```


VSOPでの実行結果　（3）



問題：与えられた k 以上を満たし，かつ maxcover, mincover がパレート最適解であること．

課題：データが追加されるとどうなるか？

ID	Place	Sex	Old	Hobby
A	Tokyo	Male	35	Sports
B	Tokyo	Female	38	Arts
C	Yokohama	Male	36	Arts
D	Yokohama	Female	32	Sports
E	Tokyo	Male	33	Sports
F	Tokyo	Female	35	Sports
G	Yokohama	Male	32	Arts
H	Yokohama	Female	31	Arts

◆単調な特殊化にはならない．
◆一般化される場合はそのまま公開しても問題は生じない．
◆特殊化される場合は，更新されたDBをそのまま公開できない（といわれている）
　（上の例では特殊化の方向に向かう．性別と趣味で k=2 を満たすため）

今後の課題

- 再計算時における計算量の見積もりと比較
- データのメンテナンス性に対して，VSOPの有用性を確認
- VSOPを利用した今後のプラン
  - ◆変数順序の最適化
  - ◆データの更新（追加）
  - ◆VSOPを利用した探索手法の検討