クラスタリングに基づく個人情報データベースの匿名化手法

沼尾研究室 http://www.nm.cs.uec.ac.jp

背景

近年,企業は大量の顧客データを所有し,サービスを有効に利用できるように努めている．しかし,これらのデータには多くの個人情報（例えば氏名や,生年月日,住所,メールアドレス等）が含まれるため,非常に利用価値が高い反面，個人のプライバシーを侵害する危険を孕んでいる．このような,利用価値と個人情報の保護を両立するため,データ内の情報と個人を結び付けられない「***k***-匿名化」が手法の一つとして近年注目されている．

データベースの***k***-匿名化

「***k***-匿名化」とは,テーブルの中で個人を特定できないよう，***k***人以上が同じ準識別子を持つようなデータへ元のデータを変換することである．表 1 を 3-匿名化した結果の例を表 2 に示す．表 2 は年齢を数値の範囲へ置き換え，地域のようなカテゴリ値をより大きな 区分へ置き換える一般化手法によって ***k***-匿名化されている．3 人が同じ準識別子を持つから,準識別子を元に,個人を特定することはできない．

識別子　　準識別子

属性

ID	性別	地域	年齢
A0001	男	奈良県	55
A0002	女	宮城県	52
A0003	女	兵庫県	46
A0004	女	大阪府	22
A0005	男	山形県	46
A0006	男	秋田県	36

k＝3 のとき

ID	性別	地域	年齢
-	女	関西	[22-55]
-	男	東北	[36-52]
-	女	関西	[22-55]
-	女	関西	[22-55]
-	男	東北	[36-52]
-	男	東北	[36-52]

クラスタリングに基づく***k***-匿名化

本研究では,クラスタリング手法に基づいてテーブルのレコードの集合を***k***レコード以上から成るグループへ情報損失を最小化するように分割し,更に準識別子の変換を行うことを目的している．

距離関数の定義

本研究では,数値型用、カテゴリ型用の距離関数をそれぞれ定義し,更にユークリッド距離を用いてレコード間の距離を計算する．

a. 数値型の距離関数

$$X_i = \frac{x_i - x_{ver}}{\text{標準偏差}} \qquad (1)$$

b. カテゴリ型の距離関数

$$\tau_c(v_1, v_2) = \begin{cases} 1 \\ 0 \end{cases} \qquad (2)$$

c. 情報損失関数

$$IL_{(p)} = |e| \cdot \left(\sum_{i=1}^{m} \frac{N_{i_{max}} - N_{i_{min}}}{T_{N_{i_{max}}} - T_{N_{i_{min}}}} + \sum_{i=1}^{p} \frac{H(\wedge(\cup c_i))}{H(T_{c_i})} \right) \qquad (3)$$

アルゴリズム

k-means を参考にし、2 つの段階に分かれる

第一段階： $\left\lfloor \frac{n}{k} \right\rfloor$個の初期値を取り、クラスタとして作って、各クラスタの中心を更新することでクラスタ内のサイズを***k***より多いようにする．

第二段階： 第一段階で残ったレコードへの割り当て。この時情報損失に考え合わせる。残ったレコードをクラスタに入れる時に、最小の情報損失であるクラスタを見つけてレコードを入れる。

実験

今回は属性に年齢や,性別,地域を用意し,5000 件のレコードを持つテーブルを使った．

```
==========5==========
A0004   20代   22   女   愛知県   未婚   学生
A2501   20代   23   女   愛知県   未婚   パート・アルバイト
A3188   20代   23   女   愛知県   未婚   学生
A4422   20代   24   女   愛知県   未婚   学生
A4864   20代   23   女   愛知県   未婚   学生
==========5==========
A3198   40代   48   女   新潟県   既婚   専業主婦
A3647   40代   47   女   新潟県   既婚   無職
A4191   40代   49   女   新潟県   既婚   パート・アルバイト
A4607   40代   47   女   新潟県   既婚   無職
A4874   40代   48   女   新潟県   既婚   専業主婦
==========5==========
```