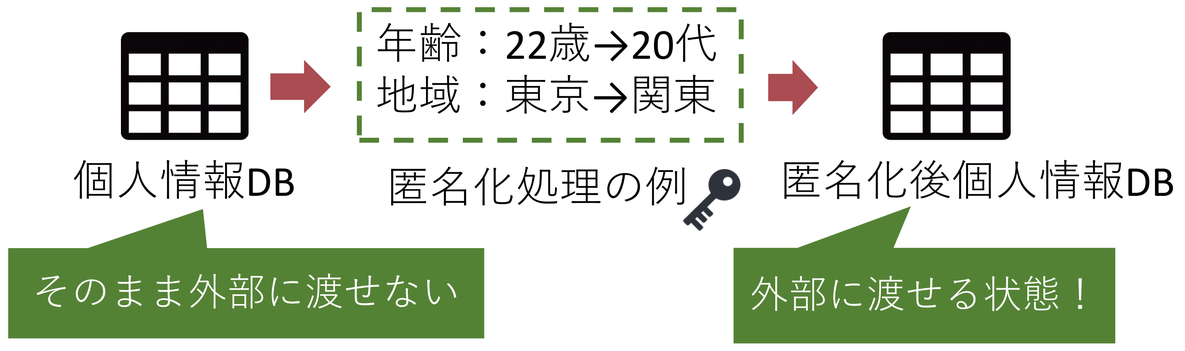
クラスタリングを用いた個人情報の匿名化

沼尾研究室　碓井愛子

研究の背景

●個人情報データベース(DB)
個人情報DBからデータマイニングを応用したサービスがあるが,個人情報は直接利用できないので匿名化の必要がある.
●目的
本研究では,k-匿名化のための個人情報DBのクラスタリング手法を提案する.



クラスタリングを用いたk-匿名化

類似ではないデータを匿名化することで,匿名化後の情報の損失が大きくなる.匿名化前にクラスタリングを行うことであらかじめ類似のデータ同士でまとめておく.

名前	性別	地域	年齢
A	女	岩手	10
B	男	沖縄	50
C	女	千葉	80

類似でないデータセット

匿名化後 →

名前	性別	地域	年齢
A	人	日本	10-80
B	人	日本	10-80
C	人	日本	10-80

匿名化後の情報の損失が大きい

名前	性別	地域	年齢
A	女	秋田	25
B	女	宮城	20
C	女	岩手	26

類似のデータセット

匿名化後 →

名前	性別	地域	年齢
A	女	東北	20代
B	女	東北	20代
C	女	東北	20代

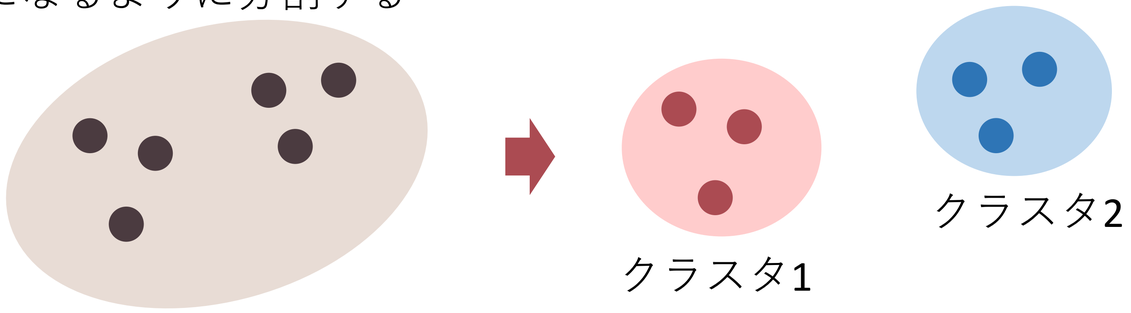
匿名化後の情報の損失が小さい

情報損失が大きいデータは統計を行う際に意味を持たない.
類似のデータでまとめる方法として,今回はm-meansという手法を用いた.mは以下の式で置く.

$$m = \frac{N}{k}$$

N: データの数
k: k-匿名化で指定したk
（同じデータが何個以上あるようにするか）

●クラスタリング
類似のデータを集めて同じかたまり（クラスタ）にする手法
●m-means
クラスタリング手法のうちの一つ
重心点とクラスタ内のデータの間の距離の二乗の総和が最小になるように分割する



●k-匿名化・・・同じデータの属性の組み合わせが少なくともk個以上ある状態にする

名前	性別	地域	年齢
A	女	奈良	25
B	男	宮城	52
C	女	兵庫	26
D	女	大阪	22
E	男	山形	56
F	男	秋田	58

k=3の場合 →

名前	性別	地域	年齢	
-	女	関西	20代	●
-	男	東北	50代	●
-	女	関西	20代	●
-	女	関西	20代	●
-	男	東北	50代	●
-	男	東北	50代	●

最低でも3つが同じデータになっている

カテゴリ型に対応したデータ変換

ID	性別	出身	年齢	婚姻	職業
A0001	男	北海道	55	未婚	自営業
A0002	女	秋田県	52	既婚	会社員
A0003	男	兵庫県	46	既婚	会社員
…	…	…	…	…	…

年齢などの数値型表記はカテゴリ型と距離を合わせるために正規化

ID	男	女	北海道	秋田県	…	自営業	会社員	…	年齢
A0001	1	0	1	0	…	1	0	…	0.56
A0002	0	1	0	1	…	0	1	…	0.52
A0003	1	0	0	0	…	0	1	…	0.44
…	…	…	…	…	…	…	…	…	…

Rでm-meansを利用するには数値型データを渡すため
性別,出身,婚姻,職業のカテゴリ型データを解析対象にある全てのデータ分の次元分を取得し,ビットベクター表現にした.

通常では扱えない文字間の距離の計算が可能になった

実験

統計解析システムRを用いて
m-meansクラスタリングを行った

ID	性別	出身	年齢	婚姻	職業
A0001	男	北海道	55	未婚	自営業
A0002	男	秋田県	52	既婚	会社員
A0003	男	兵庫県	46	既婚	会社員
…	…	…	…	…	…

解析対象の個人データ5000件

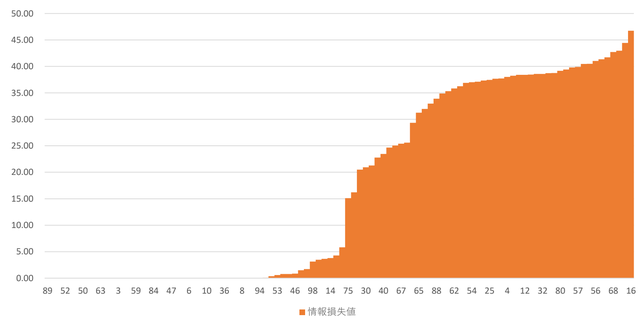
m=100

```
----------クラスタ1----------
A0095,男,大阪府,36 ,未婚,自営業
A0130,男,大阪府,36 ,未婚,自営業
A0182,男,大阪府,36 ,未婚,自営業
…,…,…,…,…,…
----------クラスタ50----------
A0044,男,大阪府,61 ,既婚,会社員
A0304,男,大阪府,55 ,既婚,会社員
A0436,男,大阪府,54 ,既婚,会社員
…,…,…,…,…,…
```

各データのクラスタリング結果

今後の課題

各クラスタの情報損失量は右図であり,各クラスタ同士の情報損失に差がある.
その理由として以下が挙げられる.
●カテゴリ型の重みづけが全て同じ重みになっている.
●各クラスタの情報損失量に緩急がある.
●クラスタ内の最低限データ数を満たしていない.
●数値型とカテゴリ型の属性間の距離関数の定義.
今後はこれらの課題点の解決を目指す.



各クラスタの情報損失量