K412 菊池研・斎藤研合同発表会 2017年2月4日

商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案

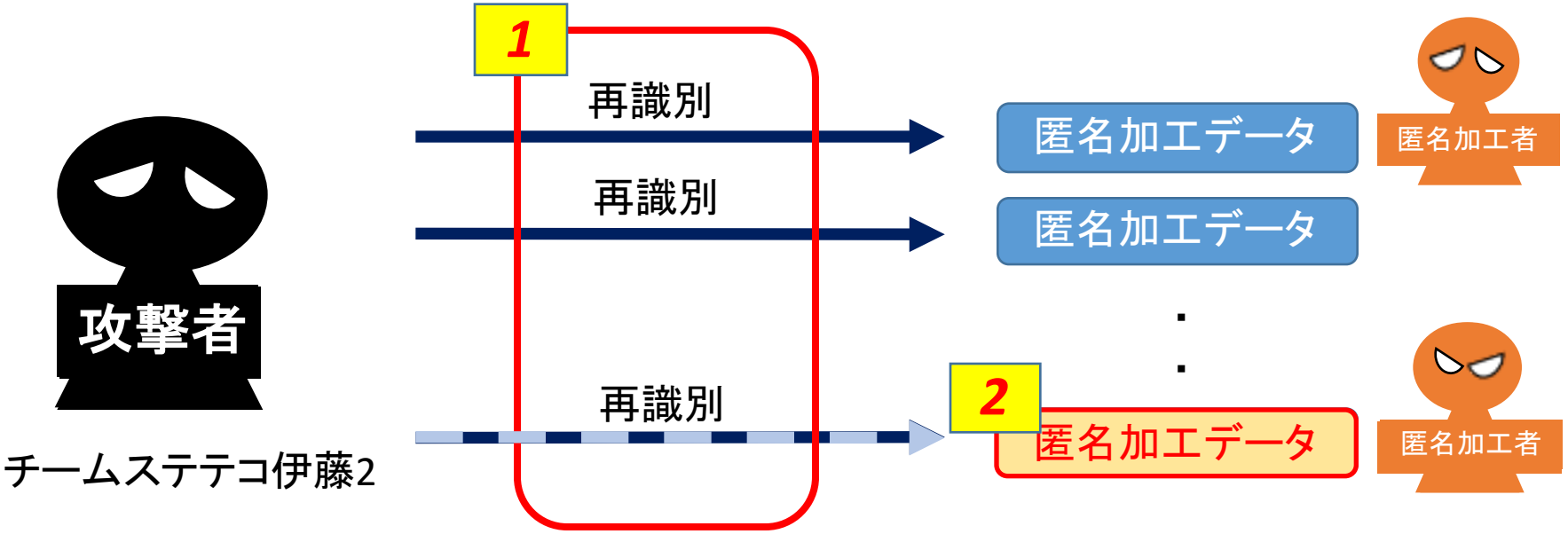
明治大学総合数理学部

菊池研4年　原田 玲央

背景

- 2016年10月，第二回「匿名加工・再識別コンテスト」が開催
- 購買履歴 = 購入商品の特徴から識別されるリスクがある

目的



チームステテコ伊藤2

1. 再識別によるリスク評価
2. 1に対する匿名加工手法の提案

Jaccard再識別アルゴリズム

Jaccard再識別

顧客が購入した商品リストについて,

Jaccard類似度が近いレコードについて再識別を行う.

元データ

行	顧客ID	商品の集合
1	u_1	$\{g_1, g_2, g_5\}$
2	u_2	$\{g_1, g_5, g_6\}$

$^2/_5$(近い)

$^1/_6$(遠い)

加工データ

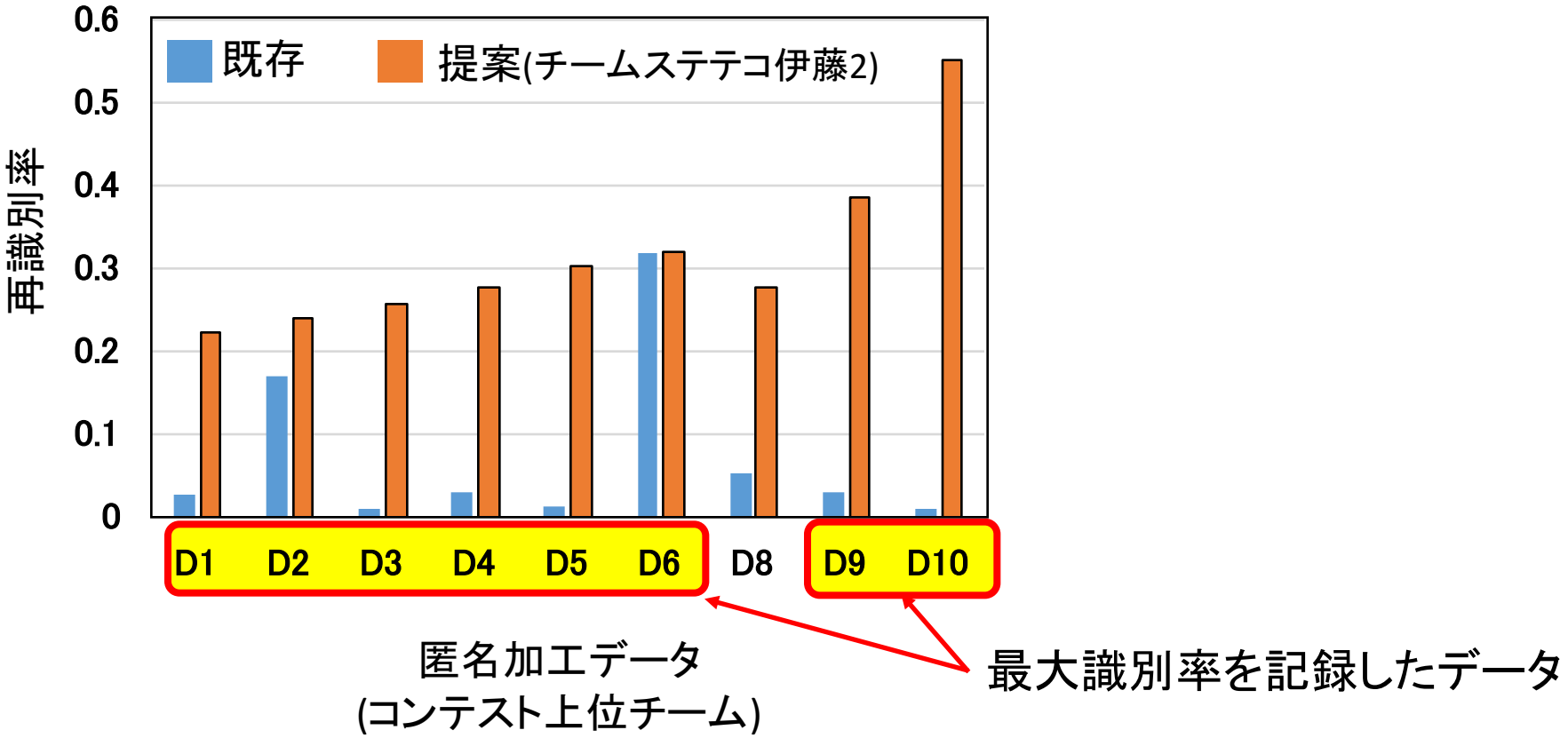
P	仮ID	商品の集合
1	u'_1	$\{g_1, g_2, g_3, g_4\}$

Jaccard類似度

2人が購入している商品のうち共通している商品の割合

例) $$J(u'_1, u_1) = \frac{|\{g_1, g_2\}|}{|\{g_1, g_2, g_3, g_4, g_5\}|} = \frac{2}{5}$$

提案アルゴリズムの再識別率



提案アルゴリズムがほとんどのデータで最大識別率を記録

購買履歴データは，商品の特徴による特定リスクが存在

耐Jaccard匿名加工手法

元データ

顧客ID	購入商品
u_1	g_1, g_2
u_2	g_3

識別不能

加工データ

仮ID	購入商品
u'_1	g_1, g_2, g_3
u'_2	g_1, g_2, g_3

統一する顧客のグループをどうやって作るか?

仮ID	購入商品
u_1	g_1
u_1	g_2
u_2	g_3

レコード追加

仮ID	購入商品
u'_1	g_1
u'_1	g_2
u'_2	g_3
u'_1	g_3
u'_2	g_1
u'_2	g_2

m

擬似レコード数
Δm

分類における問題点

コンテストのデータセットは,

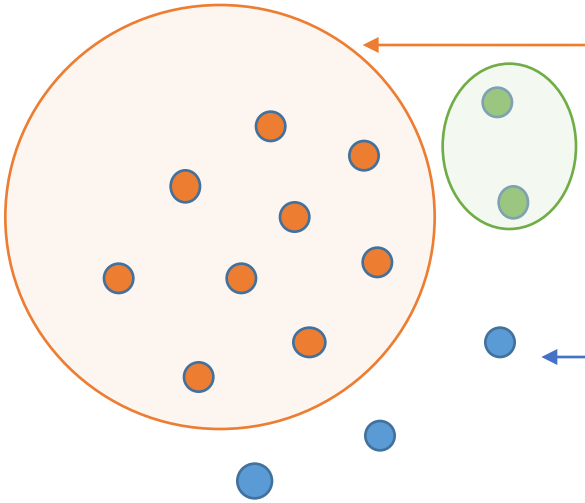
- 高次元データ
- 類似する顧客が少ない

高次元

	g_1	g_2	g_3	…	g_{2781}
u_1	1	1	0	…	0
u_2	1	0	1	…	1
u_3	0	0	0	…	1

似ていない

ナイーブなクラスタリングの例



問題点1:「巨大なクラスタ」

⇒ Δmが膨大になる

問題点2:「独立したクラスタ」

⇒ 一意に再識別される

解決方法

方式1: TF-IDF

	g_1	g_2	g_3
u_1	1	1	0
u_2	1	0	1
u_3	0	1	1
u_4	0	1	0

TF
顧客内における商品の出現頻度

IDF
商品の希少性

TF-IDF

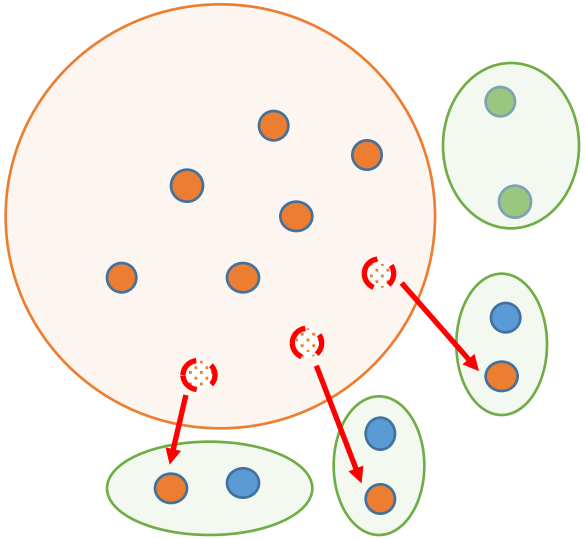
	g_1	g_2	g_3
u_1	0.7	0.6	0
u_2	0.7	0	0.7
u_3	0	0.6	0.7
u_4	0	1.1	0

TF-IDF:
どの商品が顧客を特徴づけているか

k-means

u_1	u_2	u_3	u_4

方式2: 最大クラスタの分配



識別リスクのある顧客の調整

評価

- TF-IDFの効果(方式1)
- 最大クラスタ分配の効果(方式2)
- 有用性
- 安全性
- 最適クラスタ数

  - 実験データ

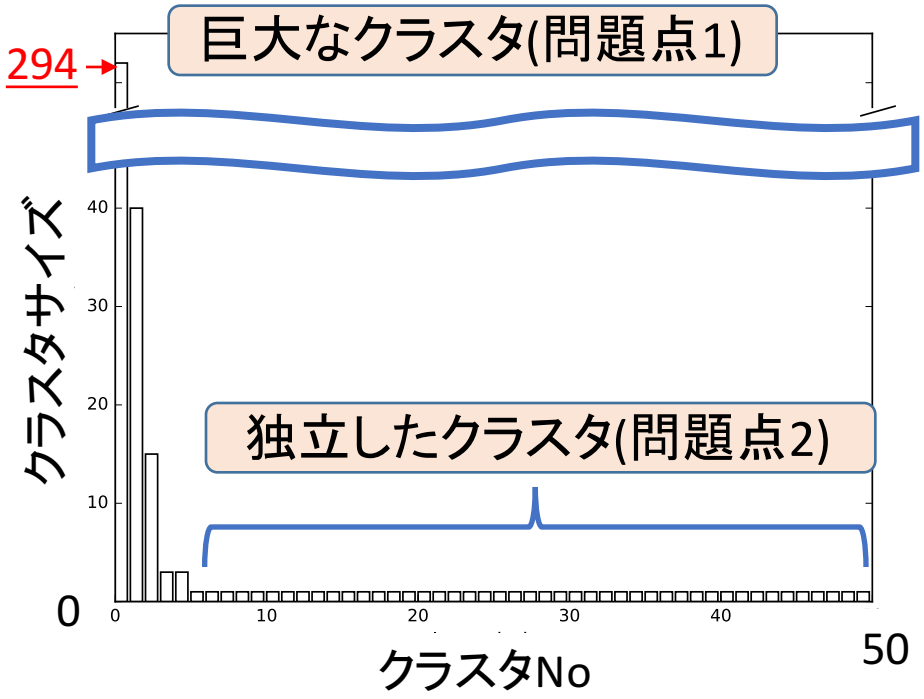
    - 顧客数$n = 400$, レコード数$\mathrm{m} = 38067$のコンテストで使われたデータセットを使用

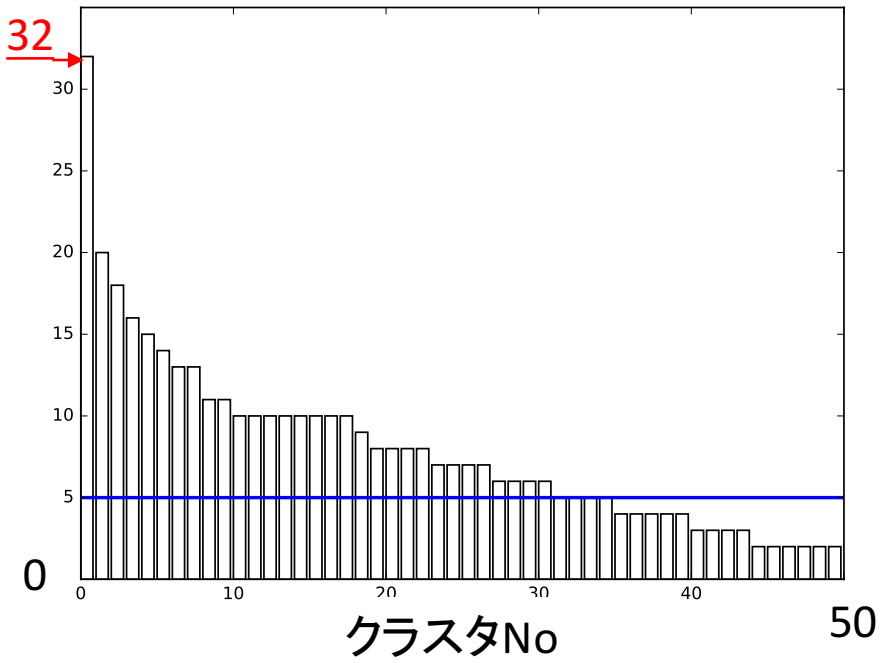
TF-IDFによる効果 (方式1)

顧客n=400人を50のクラスタに分類

(問題点)
高次元データのクラスタリング

(方式1)
TF-IDFを使ったクラスタリング





クラスタサイズの偏りを改善

最適クラスタ数

有用性

$$E(\Delta m) = -\frac{hn^3}{2c^2} + (b + \frac{h}{2})\frac{n^2}{c} - bn$$

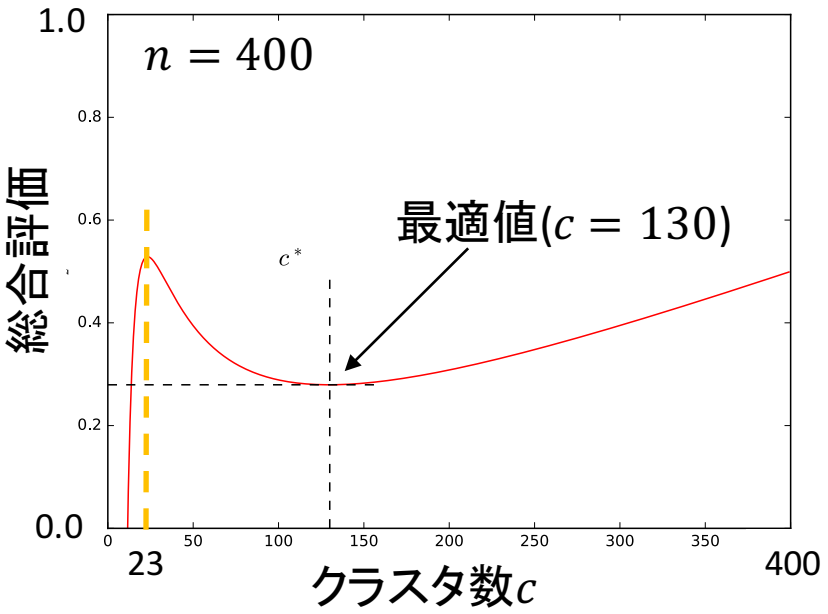
データセットの統計量
- b:平均購買商品の種類数
- h:ある2顧客間が共通して購入した商品数
- n:顧客数

安全性

$$E(Reid) = \frac{c}{n}$$

有用性と安全性はトレードオフの関係

総合評価指標: $\frac{\text{有用性+安全性}}{2}$



データセットの統計量から本手法適用による最適値を定義

まとめと今後の課題

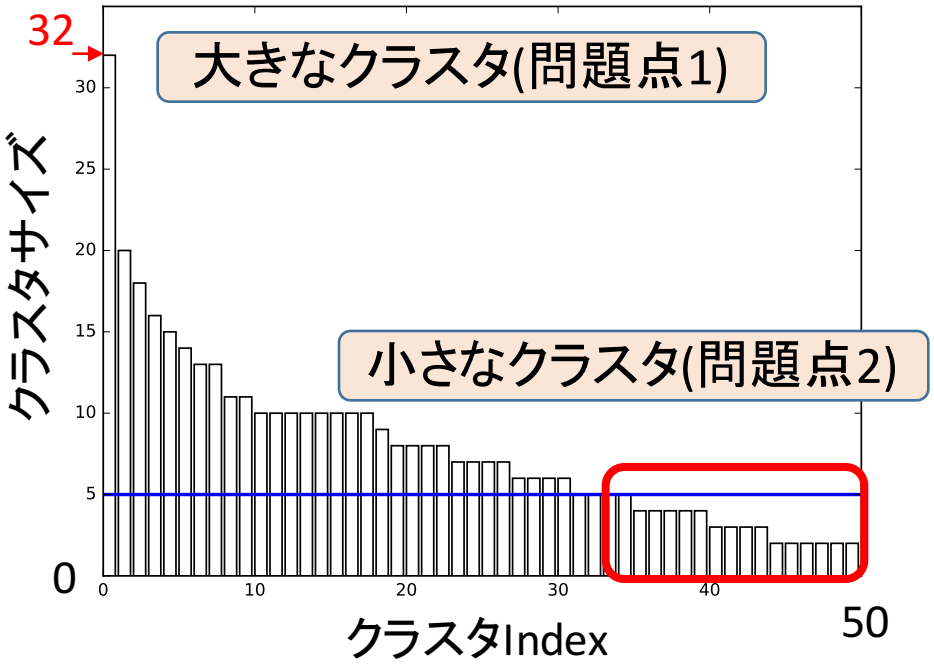
- まとめ
  - 購買履歴データに対する有効な再識別手法
  - TF-IDFを利用した匿名加工手法の提案
  - TF-IDFによる分類の評価
  - 本加工手法による理論的な最適値の定義

- 今後の課題
  - クラスタリングの精度評価
  - レコード削除や書き換えによる手法

クラスタの分配による効果(方式2)

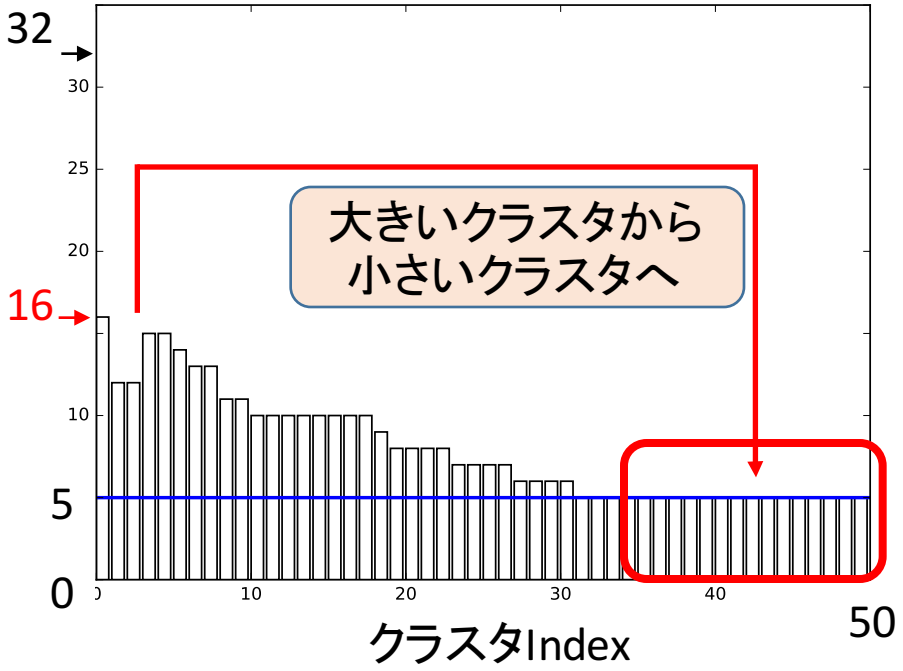
顧客n=400人を50のクラスタに分類，下限値5にした例

- 方式1



$\Delta m = 182{,}897$

- 方式2

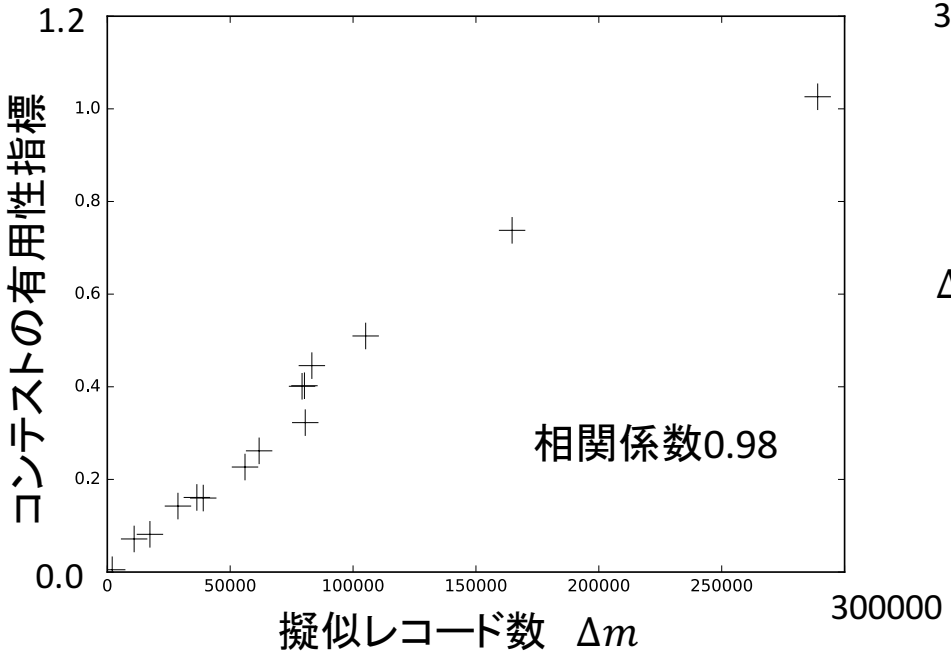


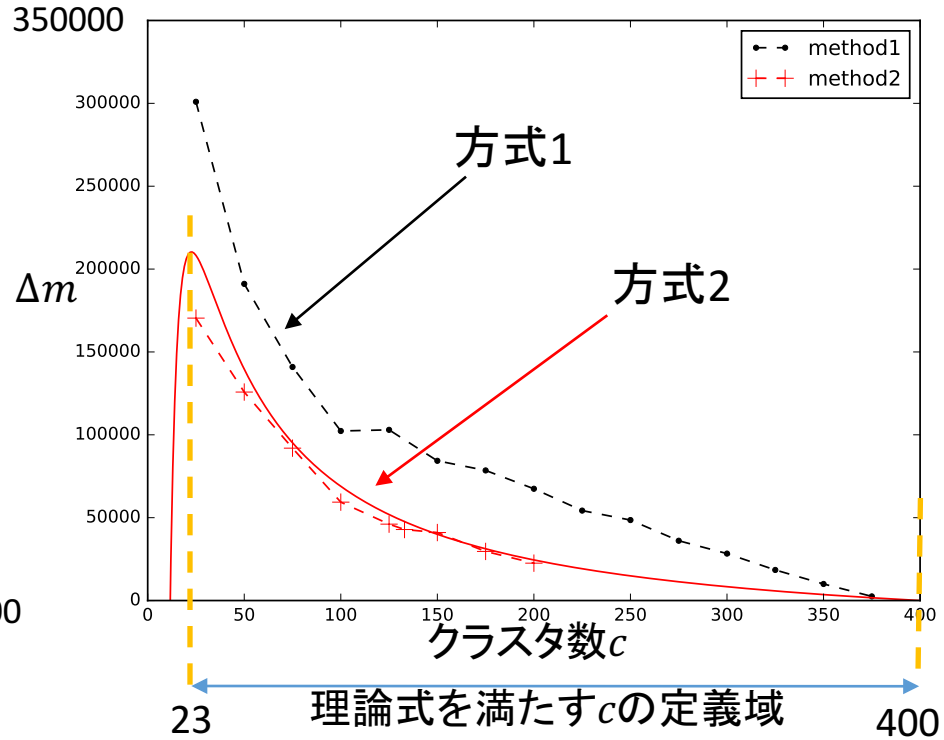
$\Delta m = 145{,}747$

独立したクラスタの解消

擬似レコード数の削減

有用性





有用性はクラスタ数に依存

データセットの統計量

- b:平均購買商品の種類数
- h:ある2顧客間が共通して購入した商品数
- n:顧客数

$$E(\Delta m) = -\frac{hn^3}{2c^2} + \underbrace{(b + \frac{h}{2})\frac{n^2}{c}}_{\text{支配項}} - bn$$

安全性

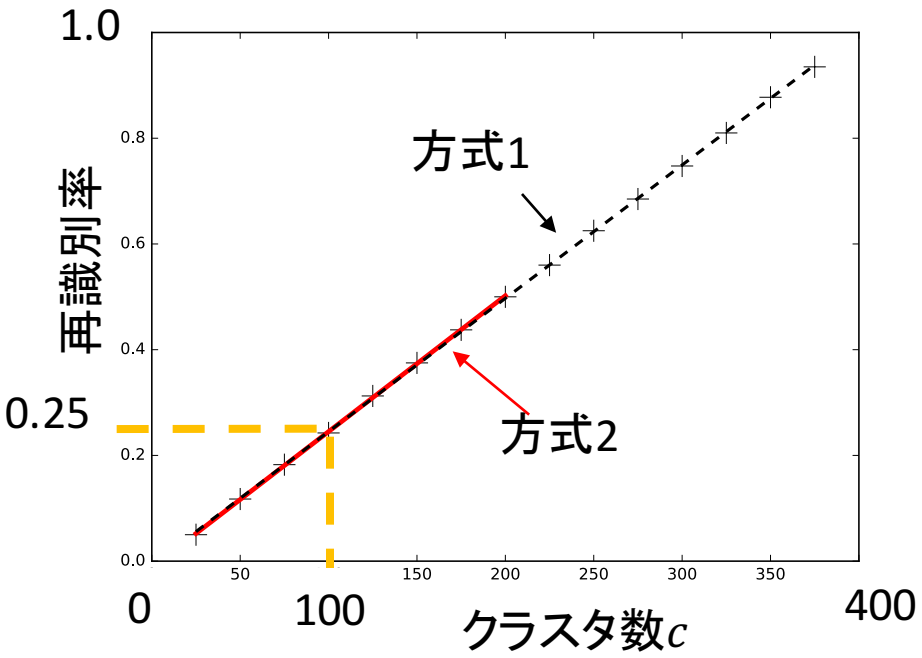
元データ

ID	商品の集合
u_1	$\{g_1, g_2\}$
u_2	$\{g_1, g_3, g_4\}$
u_3	$\{g_2, g_3\}$

Jaccard再識別

加工データ

仮ID	商品の集合
u'_1	$\{g_1, g_2, g_3, g_4\}$
u'_2	$\{g_1, g_2, g_3, g_4\}$
u'_3	$\{g_1, g_2, g_3, g_4\}$



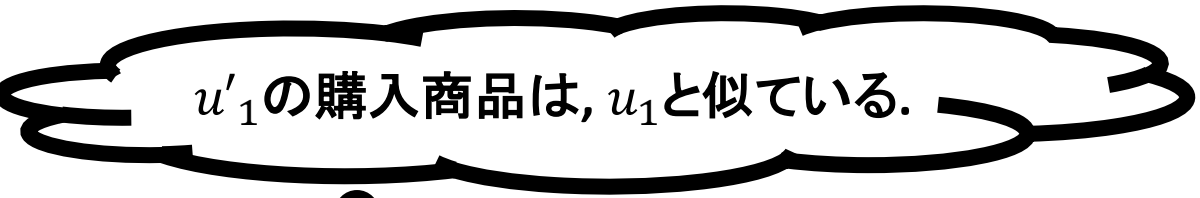
期待値 $E(Reid) = {}^{c}/_{n}$

n:顧客数, c:クラスタ数

各クラスタで一人が識別

目的

Jaccard再識別に対して有効な匿名加工手法の提案



加工データ

仮ID	購入商品
u'_1	$\{g_1, g_2, g_3\}$
u'_2	$\{g_1, g_3\}$
u'_3	$\{g_5\}$
u'_4	$\{g_4, g_6\}$

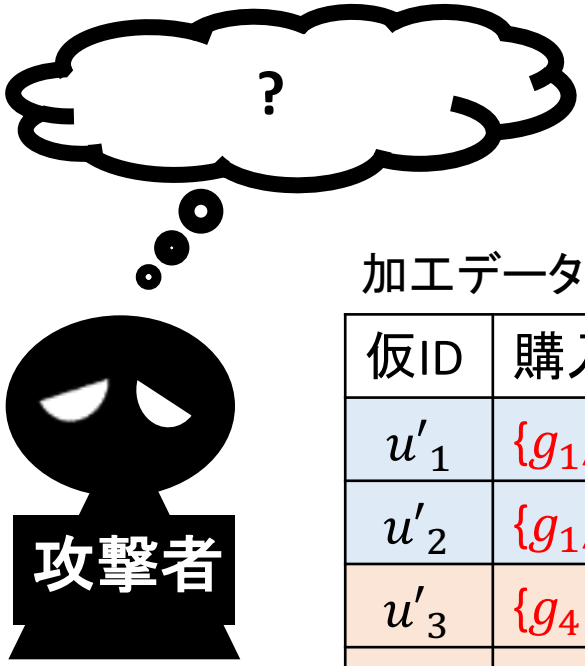
類似

元データ

顧客ID	購入商品
u_1	$\{g_1, g_2\}$
u_2	$\{g_1, g_3, g_4\}$
u_3	$\{g_5, g_6\}$
u_4	$\{g_4, g_5, g_6\}$

目的

Jaccard再識別に対して有効な匿名加工手法の提案



加工データ

仮ID	購入商品
u'_1	$\{g_1, g_2, g_3, g_4\}$
u'_2	$\{g_1, g_2, g_3, g_4\}$
u'_3	$\{g_4, g_5, g_6\}$
u'_4	$\{g_4, g_5, g_6\}$

元データ

顧客ID	購入商品
u_1	$\{g_1, g_2\}$
u_2	$\{g_1, g_3, g_4\}$
u_3	$\{g_5, g_6\}$
u_4	$\{g_4, g_5, g_6\}$

安全性

- クラスタ内の最大要素数の顧客=識別される
- 各クラスタに1人は識別 →: 再識別率の期待値$= \frac{c}{n}$
- クラスタ内の最大要素数が複数ある場合を除く

元データ

行	顧客ID	商品の集合
1	u_1	$\{g_1, g_2\}$
2	u_2	$\{g_1, g_3, g_4\}$
3	u_3	$\{g_2, g_3\}$
4	u_4	$\{g_2, g_5\}$
5	u_5	$\{g_4, g_5\}$

加工データ

P	仮ID	商品の集合
1	u'_1	$\{g_1, g_2, g_3, g_4\}$
2	u'_2	$\{g_1, g_2, g_3, g_4\}$
3	u'_3	$\{g_1, g_2, g_3, g_4\}$
5	u'_5	$\{g_2, g_4, g_5\}$
4	u'_4	$\{g_2, g_4, g_5\}$

推定行番号

Q
2
2
2
4 or 5
4 or 5