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

原田 玲央 †

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室 †

1 はじめに

2015 年 9 月の個人情報保護法の改正により, 匿名加工情報という新たな枠組みが定義された. それに伴い, 2016 年 10 月に安全で有用性の高い匿名加工技術の開発促進を目的に, 第二回匿名加工・再識別コンテスト PWSCUP2016 が開催された [1].

著者の属するグループは, 本コンテストに参加し, 最も有用性の高い匿名加工データを最も正確に再識別した. 我々は, 顧客ごとの購入商品の集合に固有の特徴を有していることに注目し, 商品集合による再識別アルゴリズムを導入した. 本稿では, そのアルゴリズムを述べ, それによる再識別リスクを明らかにする. この商品集合による再識別の問題の対策として, 購入商品が類似している顧客を次のように加工する 2 つの手法を提案する.

1. 商品集合ベクトルの TF-IDF によるクラスタリング
2. 最小クラスタサイズを制約する新アルゴリズム

これまでに, クラスタサイズを制約する手法として, 全てのクラスタサイズを均一にするまでクラスタを二分割していく手法 [2] が緒方らによって提案されている. これに対して, 均一ではなく, 最小クラスタサイズのみを設ける点が新規である.

提案手法によってクラスタリングされた顧客の購入商品を統一するように疑似データを追加し, その有用性と安全性について評価する.

2 再識別

2.1 データセットの特性

PWSCUP2016 では共通データセットとして, Online Retail Dataset[3] が使用された. 本データセットは, 英国のオンライン店舗において 2010 年 12 月から約 1 年間に渡り実際に取引された購買履歴データで, UCI Machine Learning Repository*から公開されている.

本データセットにおいて,

- $U = \{u_1, \ldots, u_n\}$: 顧客の集合
- $I(U) = \{g_1, \ldots, g_\ell\}$: 全顧客が購入した商品の集合
- $I(u_i) \subseteq I(U)$: 顧客 u_i が購入した商品の集合
- b: 一人あたりの年間平均購買商品種類の数

を定義し, 加工データの顧客は U' とする. jaccard 値は,

$$J(A, B) = \frac{|I(A) \cap I(B)|}{|I(A) \cup I(B)|}$$

で定まる集合 A, B 間の類似度である.

顧客 n 人から全ての異なる 2 人の組み合わせにおける jaccard の平均値を μ とする. また, 2 顧客が購入した商品集合の積の大きさを $h = |I(u_i) \cap I(u_j)|$ と定める. b, μ を使って,

$$h = \frac{2b\mu}{b + \mu}$$

と表すことができる.

本データセットの特性値は, $n = 400, m = 38087, b = 65, h = 4, \mu = 0.03$ である.

2.2 jaccard 再識別アルゴリズム

本コンテストにおいて, 匿名加工者は, 個人情報である顧客マスターデータ M と各顧客の購買取引の履歴を表すトランザクション T を加工して M', T' を作成し, M におけるインデックス u と M' における u' の行置換を表す行番号 P を提出する. 再識別者は, 元データ M,T を頼りに, 加工された M',T' を解析して, 推定行番号 Q を導出する. Q と P を比較することにより再識別率を定める.

そこで我々は, 商品集合の特徴量をもとにして特定を行う識別手法について考える. 元データと加工データのそれぞれについて過去に購入した商品リストを顧客ごとに算出し, 集合の類似度を示す jaccard 係数を用いて最も

†Reo Harada, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

*https://archive.ics.uci.edu/ml/datasets/Online+Retail

表 1　商品の特徴による再識別リスク

加工データ	最大再識別率 (a)	jaccard 再識別 (b)
D_1	0.2225	*0.2225
D_2	0.2375	*0.2375
D_3	0.2550	*0.2550
D_4	0.2750	*0.2750
D_5	0.3025	*0.3025
D_6	0.3175	*0.3175
D_8	0.3725	0.2750
D_9	0.3850	*0.3850
D_{10}	0.5500	*0.5500

近い顧客同士を結びつける. 本 jaccard 再識別アルゴリズムを Algorithm 1 に示す.

Algorithm 1 jaccard 再識別

Input: M, T, M', T'

Step 1.

元データ M, T と加工データ M', T' について顧客ごとに購入した商品集合を各々 $I(u_i), I(u'_i)$ $(i = 1, \ldots, n)$ とする.

Step 2.

加工データの顧客 $j = 1, \ldots, n'$ について, jaccard 類似度が最大である元データの顧客

$$i^*_j = \underset{i \in \{1,\ldots,n\}}{\arg\max}\, J(I(u'_j), I(u_i))$$

と定める.

Output: 選択した顧客の行番号列 $Q = (i^*_1, i^*_2, \ldots, i^*_n)$ を返す.

2.3　評価結果

PWSCUP2016 の本戦に参加した自チームを除く上位 9 チームから提出された購買履歴データを匿名加工したデータを $D_1, \ldots, D_6, D_8, \ldots, D_{10}$ とする.

表 1 に評価結果を示す. (a) 列はコンテストで最も高いチームの識別率, (b) 列は本アルゴリズムによるものである. *が付いている数値は, 提案 jaccard 識別手法が加工データに対して最も再識別率が高かったことを表す. コンテストのルールに則ると, 最も優秀な加工データ D_1 でも 22.25% の顧客が再識別されている [4].

3　提案加工手法

3.1　jaccard 再識別の対策

本節では, jaccard 再識別手法の対策として, レコード置換による匿名化やレコード値を変更する加工をせず,

(b)

	U	$I(U)$
1	u_1	$\{g_1, g_2\}$
2	u_2	$\{g_1, g_3, g_5\}$
3	u_3	$\{g_4, g_5\}$

(c)

P	U'	$I(U')$
1	u'_1	$\{g_1, g_2, g_3, g_4, g_5\}$
2	u'_2	$\{g_1, g_2, g_3, g_4, g_5\}$
3	u'_3	$\{g_1, g_2, g_3, g_4, g_5\}$

(a) T

U	伝票ID	商品
u_1	100	g_1
u_1	300	g_2
u_2	500	g_1
u_2	500	g_3
u_2	600	g_5
u_3	...	...

(d) T'

U'	伝票ID	商品	
u'_1	100	g_1	
u'_1	300	g_2	
u'_1	300	g_3	*
u'_1	300	g_4	*
u'_1	300	g_5	*
u'_2	...	...	

＊疑似レコード

図 1　疑似レコードの追加方法

疑似レコードを追加するだけの匿名加工手法について検討し, 個人特定リスクの一つである商品集合の特徴量を顧客間で統一する加工を試みる.

疑似レコードの追加アルゴリズムの動作例を図 1 に示す. 元データ M, T について顧客ごとの購入商品を集計する (a)(b).

顧客 u_1, u_2, u_3 の購入した商品集合を

$$I(u'_1) = I(u'_2) = I(u'_3) = I(u_1) \cup I(u_2) \cup I(u_3) = \{g_1, g_2, g_3, g_4, g_5\}$$

と共通にする (c). 例えば, u_1 の仮 ID に対応する u'_1 に商品 $\{g_3, g_4, g_5\}$ を新たな疑似レコードとして各顧客の適当な伝票 ID に追加する (d).

しかし, 全員が同じ商品を購入したとすると変更が大きすぎるので, 購入商品が類似している顧客をクラスタリングし, 各クラスタ毎に購入商品を統一する. クラスタ内の商品集合を統一することで jaccard 再識別による個人の特定は, 元データの商品集合の要素数が最多な顧客のみに限定することができる.

クラスタ数を c, 顧客クラスタの集合を $X = \{x_1, ..., x_c\}$, クラスタサイズを $s_i = |x_i|$ と定義すると, クラスタ x における疑似レコード数は, $\Delta m = \sum_{u \in x} |I(x)| - |I(u)|$ である.

3.2　レコード間距離の定義

顧客ごとの商品集合のデータは高次元であり, そのままクラスタリングに適用しても意図した結果が得られない [5]. そこで, 我々は文書をクラスタリング [6] する際

Algorithm 2 TF-IDF による購入商品の重み付け

Input: 顧客 $u_i \in U$, 商品集合 $I(u_i), c$

Step 1. 顧客 u_i の全商品数 ℓ 次元の特徴ベクトルを $\boldsymbol{v}_i = (f_{i1}, f_{i2}, \ldots, f_{i\ell})$ と表す. ここで,

$$f_{ij} = \begin{cases} 1 & \text{if} \quad I(u_i) \ni g_j \\ 0 & \text{otherwise} \end{cases}$$

とする.

Step 2. ある商品 g_j を購入した全顧客の集合を $D_j = \{u_i \in U \,|\, I(u_i) \ni g_j\}$ と表す. f_{ij} の TF-IDF による重みを

$$f'_{ij} = \frac{f_{ij}}{\sum_{k=1}^{\ell} f_{ik}} (\log \frac{n}{|D_j|} + 1)$$

と定め, 重み付けした顧客 u_i の特徴ベクトルを $\boldsymbol{v}'_i = (f'_{i1}, f'_{i2}, \ldots, f'_{i\ell})$ で表す.

Step 3. 特徴ベクトル $\boldsymbol{v}'$ 間の cos 類似度を算出して顧客 U を k-means を使ってクラスタリングする.

Output: $X = \{x_1, x_2, \ldots, x_c\}$

に用いる TF-IDF を使い, 各商品に対して重み付けをしてクラスタリングを行う. Algorithm 2 に TF-IDF を用いたクラスタリングを示す.

3.3 方式 1(既存クラスタリングベース)

TF-IDF による商品を重み付けと cos 類似度を使った k-means によるクラスタリングを行い, 各クラスタ内で商品集合の和集合をとり, 疑似レコードを追加する手法を提案方式 1 とする.

大きいクラスタに属する顧客ほど, 追加すべき疑似レコード数は増える. 逆に, $s_i = 1$ のクラスタの顧客は疑似レコードを追加しないので一意に特定することができる.

3.4 方式 2(調整アルゴリズム)

方式 1 のクラスタサイズの偏りを改善するため, 全てのクラスタサイズが下限値 s_{min} を下回らないようにクラスタを調整するアルゴリズムを方式 2 を提案する. 購入商品が最も類似する顧客を, 最大クラスタ x_{max} から s_{min} 未満のクラスタへ移動し, 全てのクラスタサイズが s_{min} 以上になるよう繰り返す.

取りうるクラスタサイズ s_{min} の下限値はクラスタ数 c に依存し, その値域は $s_{min} \in \{2, 3, \ldots, \lfloor \frac{n}{c} \rfloor\}$ である.

4 評価

4.1 Δm の理論値

疑似レコード追加手法における Δm の理論値を求める. 追加レコード数 Δm の理論値は,

$$E(\Delta m) \fallingdotseq (b + \frac{h}{2}) \frac{n^2}{c} \tag{1}$$

と, データセットの特性を示す b, μ, n をパラメータとした c の式で近似することができる [7].

4.2 有用性

$n = 400$ における s_{min} と追加疑似レコード数の関係を表 2 に示す. 各 c について $s_{min} = \lfloor \frac{n}{c} \rfloor$ の時, Δm は最小をとる. また, 方式 2 を適用することによる jaccard 類似度の標準偏差は c, s_{min} の値に対して 0.01 未満を示し, 安定している. また, 方式 2 は方式 1 の追加手法に比べて, Δm を約 53% と大幅に抑えることができている.

4.3 安全性

本手法による加工データに対して jaccard 再識別を行うと, クラスタ内のどの顧客 $u' \in x$ も, 元データの商品要素数が最多な顧客 $u \in x$ に識別される. よって, 方式 1, 方式 2 ともに再識別率の期待値は

$$E(Reid) = \frac{c}{n} \tag{2}$$

である.

また, 各 c における再識別率の実測値 $Reid$ を表 2 に示す. 再識別率は, s_{min} の値によらず, c に依存する.

4.4 最適クラスタ数

データを加工すると, 一般的に有用性が悪くなり, 安全性が高くなる. しかし, この 2 つの指標を総合的に評価するにはユースケースやデータ構成に依存する. 本稿では, コンテストでの総合評価 [1] に使用された $\frac{U+E}{2}$ の U を Δm と置き換え,

$$\frac{\alpha E(\Delta m) + E(Reid)}{2} \tag{3}$$

を用いてクラスタの最適値 c^* を定める. c^* は (3) 式が極小をとるときの c の値である. ここで, α は Δm を $0 \leq E(\Delta m) \leq 1$ に正規化する係数とする. $n = 400, b = 65, \mu = 0.03$ のデータセットを方式 2 の手法に適用した時の最適値 c^* を図 2 に示す. 評価値が極小となるのは,

表 2　s_{min} に対する Δm の関係

	$c = 50$			$c = 75$			$c = 100$			$c = 125$		
	Δm	jaccard	Reid	Δm	jaccard	Reid	Δm	jaccard	Reid	Δm	jaccard	Reid
方式 1	182897	0.1728	0.1235	141696	0.2402	0.1858	128568	0.3060	0.2488	97581	0.3692	0.3120
$s_{min} = 2$	183902	0.1729	0.1223	136526	0.2403	0.1860	99228	0.3061	0.2475	60492	0.3687	0.3105
$s_{min} = 3$	175449	0.1726	0.1222	112781	0.2394	0.1855	68357	0.3041	0.2480	*46101	0.3667	0.3102
$s_{min} = 4$	162474	0.1723	0.1218	*91946	0.2382	0.1855	*59374	0.3044	0.2465			
$s_{min} = 8$	*125798	0.1681	0.1218									

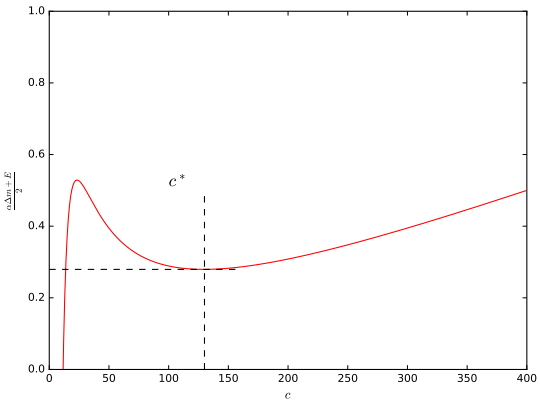

図 2　方式 2 の最適値 c^*

$c^* = 130$ の時である. ただし, 定義域は $23 \leq c \leq 400$ であることに注意したい [7].

顧客数 n についてのクラスタ数の最適値 c^* を考えよう. (1) 式と (2) 式を (3) 式に代入した極小値から最適値

$$c^* = \sqrt{\alpha(b + \frac{h}{2})n^3} \tag{4}$$

を得る. ただし, α は n に依存する変数であることに注意したい.

(3) 式と α を与えたとき, (4) 式を用いて, データセットの特性 b, μ, n から, 方式 2 における最適値 c^* を導出することができる. 例えば, $n = 4000, b = 65, \mu = 0.03$ のデータセットに対しては, $c^* = 1427$ より, 再識別率 $E(Reid) = 0.3567$, $E(\Delta m) = 490650$ が方式 2 における最適な加工である.

5　おわりに

PWSCUP2016 の結果に基づき, 購入商品の特徴を用いた再識別手法における特定リスクを明らかにした. その対策として疑似レコードを追加する匿名加工手法を提案した. 提案手法は, 商品の類似している顧客を TF-IDF による重み付けを取り入れてクラスタリングし, クラスタサイズの下限値を設けることで追加疑似レコード数を抑える. また, 提案方式における追加疑似レコード数と再識別率の理論値を求め, データセットの特性から提案加工手法の最適なパラメータを導出できることを確認した.

参考文献

[1] 菊池浩明, 小栗 秀暢, 野島 良, 濱田 浩気, 村上 隆夫, 山岡 裕司, 山口 高康, 渡辺 知恵美, “PWSCUP:履歴データを安全に加工せよ”, CSS 2016, pp. 271-278, 2016.

[2] 緒方悠人, 遠藤靖典, “K-Member Clustering 問題に関する一考察”, FSS 2013, pp. 61-66, 2013.

[3] Daqing Chen, Sai Liang Sain, and Kun Guo, “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining,” Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197-208, 2012.

[4] PWS 実行委員会, “PWSCUP 匿名加工・再識別コンテスト”, (https://pwscup.personal-data.biz), 2016 年 12 月参照.

[5] 長谷川聡, 菊池亮, 正木彰伍, 濱田浩気, “行列分解を利用した確率的 k-匿名性を満たす高次元データ公開法”, CSS 2016, pp. 936-942, 2016.

[6] Rakesh Chandra Balabantaray, Chandrali Sarma and Monica Jha, “Document Clustering using K-Means and K-Medoids”, arXiv preprint arXiv:1502.07938, 2015.

[7] 原田玲央, 伊藤聡志, 菊池浩明, “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”, SCIS 2017, (発表予定), 2017.