匿名加工データにおける一般化加工手法 k-匿名化の安全性評価，商品カテゴライズ化による安全性の向上

中村幸輝 †

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

2013 年 6 月の JR 東日本の Suica 乗降履歴データ販売案件は，ビッグデータの第三者提供における，個人を特定できないようなデータ加工の課題を考える大きな契機となった．2017 年 5 月に施行された改正個人情報保護法にて「匿名加工情報」が導入された [1].

匿名加工情報の対象は大きく二つある．一つは個人情報である．これは氏名，住所，マイナンバー，電話番号などで，情報を組み合わせると個人が識別される可能性のある情報である．これは削除，もしくは仮名化しなければならない．もう一つは，移動や購買の履歴情報である．これは単一の情報だけで個人が識別されるわけではないが，加工前データと比較することで個人が特定されてしまうリスクがある．匿名加工では，この履歴情報をいかに，有用性高く，かつ安全性の高いデータにするかが問われる．昨年，匿名加工データの有用性を RFM 分析を使って評価する研究を行なった [2].

匿名加工情報の有用性と安全性で競う大会が PWSCUP[3] である．この大会は，チームごとに購入履歴データを匿名加工し，他のチームの匿名加工データを再識別するものである．2018 年に行われた PWSCUP2018 では匿名加工の手法は一般化手法のみである．しかしながら，3,200 以上ある商品は不規則な ID が振られて管理されており，「トマト」を「野菜」などのようにカテゴリにて一般化することができるかもしれないと考えた．その安全性基準を表 1 に示す．例えば，16 人識別に挑戦して 13 人以上の識別に成功すれば安全性基準を満たしていないことになる．

そこで本研究では，全 3,253 種類の商品を家具、装飾品などのカテゴリーにカテゴライズを行うことで一般化し安全性の向上を図る．商品のカテゴリーからはユーザが何を買ったか把握できないため，匿名性が増し，安全性が向上する．安全性を k-匿名化の k を用いて $1/k$ とする．また，PWSCUP2018 の有用性評価プログラムを用いて，有用性を評価し，これに安全性を加えた結果を総合値として，最適な k として k を検討する．

†Koki Nakamura, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

表 1　PWSCUP2018 で用いられた安全性基準

識別試行人数	識別目標人数
7	7
11	10
13	11
16	13
19	15
500	308
1,000	612

2 k-匿名化プログラムについて

2.1 元データの説明

まず，匿名加工を行うデータについて説明する．本研究では,PWSCUP2018 で用いられた UCI Machine Learning Repository[4] の Online Retail Data Set (2010 年から 1 年間の英国のオンライン小売店での購買履歴，8 属性，レコード) を PWSCUP2018 用に編集したものを利用する．本データは 81776 レコード 5 属性のデータである．表 2 に本研究で使用する属性を示す．顧客 ID はランダムに割り振られた 5 桁の数値である．購入日は 2010/12/1 から 2011/11/30 までのデータである．商品 ID は数値と文字で表されるデータである．単価の単位は$で，最小値が 0.04，最大値が 2,500 のデータである．購入数量は最小値が 1，最大値が 2,880 のデータである．また，表 3 に本データの概要を示す．

表 2　購買履歴データの属性

属性名	本稿での呼称	内容
CustomerID	顧客 ID	購買をした顧客の ID（5 桁数値）
Day	購入日	購買した年月日（yyyy/mm/dd）
GoodsID	商品 ID	購買した商品の ID（数値，文字）
Price	単価	購買した商品の単価 ($)
Quantity	購入数量	商品を購買した個数

2.2 k-匿名加工アルゴリズムの説明

購買履歴データの k-匿名加工アルゴリズムを説明する．

例として本章では，3-匿名加工のアルゴリズムを説明する．

表 3　購入履歴データの例

CustomerID	Day	GoodsID	Price	Quantity
13047	2010/12/1	84879	1.69	32
15236	2011/2/14	22720	4.95	3
15514	2011/4/28	10125	0.85	1
18219	2011/5/9	23132	5.75	3
14944	2011/7/18	23345	8.5	2
16241	2011/11/30	23322	2.95	2

1. ユーザごとに分類し，購買回数で降順にソート:
　一般化では，類似するユーザをクラスタリングする必要があるため，まず，購買回数を揃える．まずは購買回数を降順でソートを行う．例えば，3-匿名加工の時，上から 3 人に同じクラスター ID を振る．ユーザごとの購買回数を表 4 に示す．表 4 から分かる通り，ユーザ 12415 は 4,289 回購入していることになる．例えば，クラスター ID が 1 のグループは，12415，12388，15005 となる．また，本研究ではユーザを k 人でグルーピングするために，1,000 を k で割った余りの人数のユーザを下から削除する．(3-匿名化では，999 人のデータを用いる．)

表 4　ユーザごとの購買回数とユーザクラスタ

ユーザー	購入回数	クラスター ID
12415	4289	1
12388	1601	1
15005	1119	1
14769	1066	2
14527	915	2
14505	799	2
⋮	⋮	⋮
15657	1	333
18084	1	333
18184	1	333

2. レコードを価格，個数で昇順にソート
　出来るだけデータの類似度の高いデータ同士で合わせて加工するため，「価格」，「個数」で昇順にソートする．ユーザ 12415 を「価格」，「個数」で昇順にソートした．このソートを全ユーザで行い，上から順にマッチングさせ，購入回数が合わない他のユーザのレコードは下から削除する．例えば，ユーザ 12415，12388，15005 の 3 人でクラスタリングを行う場合，ユーザ 12415 は 3170 レコード（4,289-1119），ユーザ 12388 は 482 レコード（4,289-1119），削除する．その後一般化加工を行う．
3. 3 人 1 組になるようにはみ出たデータを削除
　削除したデータは PWSCUP2018 のルールに則って表**??**のように「*」に変える．
4. クラスタ内で一般化加工

表 5　ユーザごとの購買履歴データ（価格個数ソート）

CustmorID	GoodsID	Price	Quantity	ClusterID
12415	84879	300	1	1
12415	22720	120	3	1
12415	10125	100	1	1
⋮	⋮	⋮	⋮	⋮
12415	23222	1	12	1
12415	25609	1	10	1

表 6　購買履歴データ（削除）

CustmorID	GoodsID	Price	Quantity
*	*	*	*

　マッチングを行ったレコードを一般化加工する．購入日，単価，数量の 3 属性は最低値 a（最も昔の日付）と最高値 b（最も最近の日付）からなる区間 $[a;b]$ で置換する．商品 ID は，全ての要素からなる集合にする．最低値と最高値，商品 ID が一致している場合，一般化しない．例として最上位のレコード 3 件を加工する．加工前のデータを表 7，一般化加工したデータを表 8 に示す．

表 7　購買回数上位 3 名のそれぞれ最上位の購買履歴レコード（加工前）

CustomerID	Day	GoodsID	Price	Quantity	ClusterID
12415	8/5	10252	300	1	1
12388	9/21	89213	180	24	1
15005	9/11	83769	250	12	1

表 8　匿名加工された購買履歴レコード（加工後）

CustomerID	Day	GoodsID	Price	Quantity	ClusterID
12415	[8/5;9/21]	{89213;10252;83769}	[180;300]	[1;24]	1
12388	[8/5;9/21]	{89213;10252;83769}	[180;300]	[1;24]	1
15005	[8/5;9/21]	{89213;10252;83769}	[180;300]	[1;24]	1

これで 3-匿名一般化加工が完了する．
この手法では，全てのグループの要素数が完全に k となる．

3　有用性評価について

2 節のアルゴリズムに従って $k = 2$〜20 について k-匿名化処理した加工データを，PWSCUP2018 で用いた有用性評価プログラムを用いて有用性を算出した．この有用性値は小さければ小さいほど元データとの誤差が少なく有用性の高いデータと言える．
k-匿名の安全性を $1/k$ とする．これは，3 章で説明した通り，完全に k 人でグルーピングされた k-匿名加工データなので，再識別をしても k 人までしか絞れないため，安全性値を $1/k$ としている．そのため，安全性値が小さければ小さいほど安全性の高いデータと言える．この安全性値を有用性評価値に加えた値を総合値として k-

匿名加工で最適な値を評価，検証する．評価結果を表**??**に示す．

このように $k = 2$〜20 を総合値で比べたところ，$k = 6$ が最適となった．また，総合値だけで見ると $k = 2$ と $k = 15$ や，$k = 3$ と $k = 8$ 匿名加工はほぼ同じ評価であると言える．

表 9　$k = 2$〜20-匿名化加工の評価（一部省略）

k	有用性	安全性 (1/k)	総合値
2	0.348	0.500	0.848
3	0.472	0.333	0.805
4	0.542	0.250	0.792
5	0.583	0.200	0.783
6	0.615	0.166	0.781
7	0.641	0.143	0.784
8	0.680	0.125	0.805
10	0.718	0.100	0.818
12	0.748	0.083	0.831
15	0.780	0.066	0.846
17	0.807	0.059	0.866
20	0.833	0.050	0.883

4　商品カテゴライズによる総合評価の向上

本研究で用いたデータセットの中でも特に商品 ID に着目した．購入日，単価，および数量は区間として一般化することができるが，商品 ID は商品一つ一つが明らかになっているため，集合として一般化を行なっている．この商品 ID をカテゴリーに分類することで，さらに一般化できるのではないかと考えた．しかし，商品 ID は商品のジャンルに関わらず完全にランダムな乱数で振られているため，商品 ID だけではカテゴライズすることはできない．そこで Online Retail Data Set に商品毎の情報が書かれたデータがあり，これを著者の主観の元に商品を独自のカテゴリーに分類する．例えば，「CERAMIC STRAWBERRY DESIGN MUG」や「PINK HEART SHAPE EGG FRYING PAN」は「キッチン用品」，「LARGE YELLOW BABUSHKA NOTEBOOK 」や「36 PENCILS TUBE SKULLS」は「文具」としてカテゴライズする．また，商品名だけではわからないものは「不明」，商品が固有のもの，ユニークなものは「その他」としてカテゴライズする．カテゴライズした分類とその種類数を表 10，図 1 に示す．

このようにカテゴライズしたデータを元に，商品 ID をカテゴリー名に置換しデータを加工した．例えば，表**??**を表 10 を元に加工したものを表 11 に示す．さらに，2 章で説明した匿名加工プログラムで一般化匿名加工を行ったものを表 12 に示す．

このようにカテゴライズしたデータを匿名加工プログラムで一般化匿名加工を行い，有用性評価値を比較した．表 13 に，商品カテゴライズしたデータの総合評価値の評価結果を表 14 に示す．

表 10　商品のカテゴリーとカテゴリー毎の商品数

カテゴリー名	商品数
雑貨	1,233
キッチン用品	534
装飾品	471
衣類	271
家具	201
文具	161
子供用	150
ガーデニング用品	50
生活用品	45
手芸品	36
不明	34
バス用品	22
ペット用品	11
ランドリー用品	9
応急用品	8
その他	6
化粧品	5
スポーツ用品	4
楽器	2

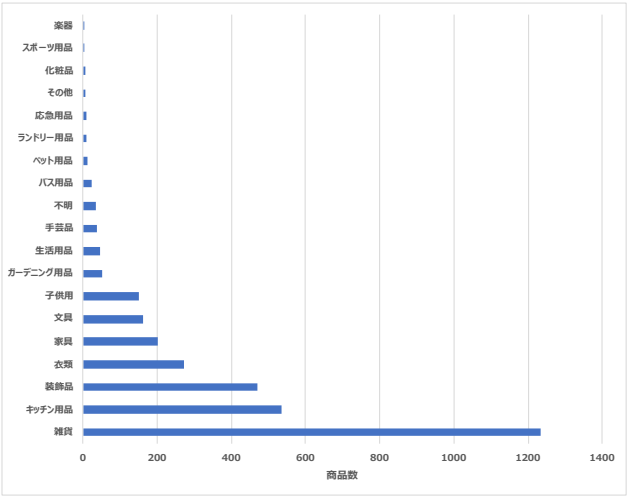

図 1　商品のカテゴリーとカテゴリー毎の商品数

表 11　商品カテゴライズ後の購買履歴レコード（加工前）

CustomerID	Day	ReceiptID	Price	Quantity	ClusterID
12415	8/5	キッチン用品	300	1	1
12388	9/21	家具	180	24	1
15005	9/11	家具	250	12	1

このように種類数が 3253 種類から 19 種類に減らすことで有用性が全体的に 0.02〜0.03 ポイント向上した．また $k = 2$〜20 の中で最適な k の値は，$k = 6$ となった．また，元データ、商品カテゴライズしたデータの匿名加工による有用性，安全性，総合評価値をグラフに示したものを図 2 に示す．

商品カテゴライズしたデータの有用性値，総合評価値

表 12　匿名加工された商品カテゴライズ後の購買履歴レコード（加工後）

CustomerID	Day	ReceiptID	Price	Quantity	ClusterID
12415	[8/5;9/21]	{ 家具; キッチン用品 }	[180;300]	[1;24]	1
12388	[8/5;9/21]	{ 家具; キッチン用品 }	[180;300]	[1;24]	1
15005	[8/5;9/21]	{ 家具; キッチン用品 }	[180;300]	[1;24]	1

表 13　$k = 2$〜20-匿名化加工の評価（一部省略）

k	元データの有用性	商品カテゴライズしたデータの有用性
2	0.348	0.324
3	0.472	0.447
4	0.542	0.517
5	0.583	0.558
6	0.615	0.590
7	0.642	0.617
8	0.680	0.655
10	0.718	0.694
12	0.748	0.725
15	0.780	0.758
17	0.807	0.785
20	0.833	0.812

表 14　商品カテゴライズしたデータの $k = 2$〜20-匿名化加工の評価（一部省略）

k	有用性	安全性 (1/k)	総合値
2	0.324	0.500	0.824
3	0.447	0.333	0.780
4	0.517	0.250	0.767
5	0.583	0.200	0.758
6	0.590	0.167	0.757
7	0.617	0.143	0.760
8	0.655	0.125	0.780
10	0.694	0.100	0.794
12	0.725	0.083	0.808
15	0.758	0.067	0.824
17	0.785	0.059	0.844
20	0.812	0.050	0.862

のグラフが示されている．図から分かる通り，元データと比べて商品カテゴライズによって 3,253 種類から 19 種類に商品の種類数を減らしたデータは 0.02〜0.03 ポイント有用性値，総合評価値共に下がる．また $k = 2$〜20 の中で最適な k の値は，$k = 6$ となった．

5　おわりに

本研究では，(1) 一般化手法における k-匿名化の $k = 2$〜20 の中で最適な値はどれか (2) 商品 ID が明確に表記されていてユーザの再識別が用意である という問題点を解決するために，(1) 有用性・安全性の二つの観点から k-匿名化の $k = 2$〜20 を評価する (2) 商品 ID を著者の主観の元独自のカテゴリーに分類することで一般化する を提案し，実装した．有用性では PWSCUP2018 の有用性評価を用い，安全性は $1/k$ で評価した結果，$k = 2$〜20 の中では 6-匿名加工が安全性，有用性の両面から見

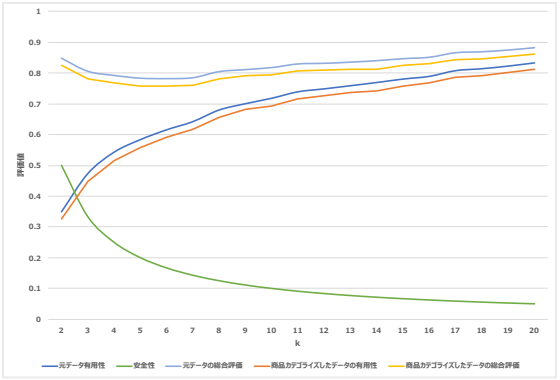

図 2　$k = 2$〜20-匿名化加工の再評価

て最適である．また，商品をカテゴライズ化したのち一般加工を行うことで安全性，有用性共に高いデータを作ることができる．有用性に関してはカテゴライズの基準によって結果が変わってくるため，マーケティングによる分析を行い元データと比べてどれほど誤差が生まれ，マーケティングに影響が生まれるか再検討する必要がある．この結果より，匿名加工情報を扱うにあたって，本研究より，さらに明確な安全性，有用性における基準，評価方法を考察する必要がある．

参考文献

[1] 小栗 秀暢，他，“匿名加工情報の作成における攻撃者知識と安全性についての一考察”，コンピュータセキュリティシンポジウム (CSS2017)，23 - 25 October 2017

[2] 小林祐貴，中村幸輝，伊藤聡志，菊池浩明，一般化匿名加工された購買履歴データの RFM 分析有用性評価，情報処理学会第 81 回全国大会，pp.3_425-3_426，2019.

[3] 濱田 浩気，他，“PWSCUP2018:匿名加工再識別コンテストの設計 履歴データの一般化・再識別 ”，コンピュータセキュリティシンポジウム (CSS2018)，pp.935-940，2018.

[4] UCI Machine Learning Repository, (`http://archive.ics.uci.edu/ml/index.php`, December 17th, 2018.)