2ZA-03

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

小林祐貴 †　中村幸輝 †　伊藤聡志 †　菊池浩明 †

明治大学総合数理学部　†

1　はじめに

匿名加工は，データから個人を特定されないようにする個人情報の加工である．2018 年 10 月に行われた匿名加工・再識別コンテスト PWSCUP2018[1] では，購買日や商品名を「一般化」する加工手法を対象として，匿名加工と再識別リスクの評価が行われた．本コンテストではパーソナルデータを含む購買履歴データについて，元データと匿名加工データとのレコード単位での平均誤差による一般的な有用性評価が行われたが，特定のユースケースにおける有用性は不確かであった．

これに対して，本研究では顧客の購買の頻度や金額などの RFM 分析のユースケースを想定し，一般化加工が行われた匿名加工データの有用性を評価する．

2　オンライン購買履歴データの分析

2.1　オンライン購買履歴データの概要

本研究では，UCI Machine Learning Repository[2] 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，541,909 レコード) のデータのうち，PWSCUP 2018 で用いられた 81,776 レコード 5 属性のデータを用いる．表 1 に購買履歴履歴データの例を示す．

表 1　購買履歴データの例

顧客 ID	購買日	商品 ID	単価	購買数量
14667	2011/11/14	21745	3.75	1
14974	2011/11/2	23392	2.08	2
14911	2011/9/30	22818	0.42	12

2.2　オンライン購買履歴データの RFM 分析

本研究では，購買履歴データの購買日，単価，数量に注目し，顧客ごとに RFM 分析を行う．RFM 分析は，R(最新購買日)，F(購買頻度)，M(購買額) の 3 つの観点で，顧客を分類し，それぞれのグループの性質を知る手法である．表 2 に元データと匿名加工データの RFM 結果の例を示す．顧客 12348 は最新日から 66 日前に最後の購買を行い，年間に 4 回，計 1797.24 ポンドの購買を行ったことを示している．

図 1 に顧客の年間購買総額と累積購買構成比率の上位 100 顧客を示す．この分析により，少数の上位顧客が全体の売上のほとんどを支えていることがわかる．

図 2 に顧客の最新購買日と年間購買頻度の分布を示す．赤線は R の 10 分位値，青線は F の 10 分位値を示している．この顧客の分布から，R が小さく F が大きい顧客は優良顧客，R も F も小さい顧客は新規顧客といった顧客判別をすることができる．

表 2　元データと匿名加工データの RFM 結果の例

	元データ			匿名加工データ		
顧客 ID	R	F	M	R	F	M
12348	66	4	1797.24	36	10	1387
12349	9	1	1757.55	32	3	1050
12354	223	1	1079.4	209	1	984.75

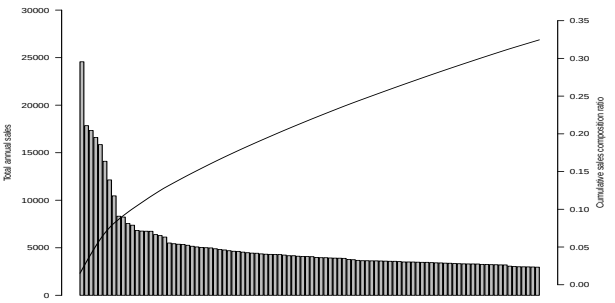

図 1　顧客ごとの年間購買総額と累積購買構成比率

3　購買履歴データの匿名加工

本研究では k-匿名化を用いる．k-匿名化は，同一属性を持つレコードを k 件以上になるように変更することで，個人が特定される確率を k 分の 1 以下に低減する．以下に本研究の匿名加工アルゴリズムを示す．

(1) レコード数で顧客をソート：ソートすることで，

RFM Analysis on Anonymized Customer Retail Records Data.

†Yuki Kobayashi, Koki Nakamura, Hiroaki Kikuchi, School of Interdisciplinary Mathematical Science, Meiji University.

†Satoshi Ito, Graduate School of Advanced Mathematical Sciences, Meiji University.

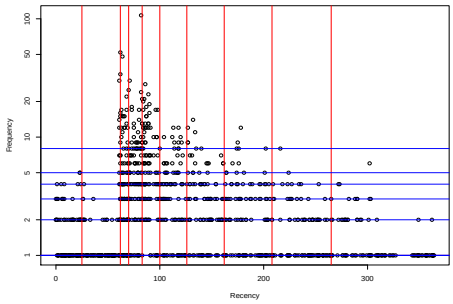

図 2　顧客の最新購買日と年間購買頻度を表す散布図

レコード数の近い顧客を探し出す．

(2) マッチング：顧客をレコード数順に k 人ずつマッチングしてクラスタとする．

(3) レコード削除：k 人のレコード数が異なる場合はレコード削除を行う．

(4) 匿名加工：PWSCUP2018 のルールに従い，一般化の匿名加工を行う．

表 3 に 2-匿名加工データの例を示す．顧客 23 と顧客 407 が 1 月 1 日から 2 月 21 日の区間のいずれかの日に，単価が 1.0 から 8.0 ポンドである商品 229 または 201 を 3 個以上 12 個以内の数購買していることを示している．

表 3　匿名加工データの例

顧客 ID	購買日	商品 ID	単価	購買数量
23	[01/01,02/21]	{229,201}	[1.0,8.0]	[3,12]
407	[01/01,02/21]	{229,201}	[1.0,8.0]	[3,12]
166	[01/01,03/06]	{225,848}	*	[1,5]
843	[01/01,03/06]	{225,848}	*	[1,5]

4　匿名加工データの有用性評価・安全性評価

4.1　有用性評価・安全性評価

本研究では，元データと匿名加工データの RFM 結果から，匿名加工データの有用性評価を行う．区間化された匿名加工データの RFM を計算する際は，一様に任意の値を選び出す作業を 100 回繰り返し，100 回分の平均値を区間開始値に足しあわせた値を使用する．ただし，同じ区間の購買日に匿名加工されている場合は，同じ購買日として扱う．

表 4 に $k = 2, 3, 4$ の匿名加工データの有用性結果と安全性結果を示す．RFM の有用性は，R，F，M のランクから計 1000 ランクにクラス分けし，元データと匿名加工データの顧客のクラスが一致した割合で評価する．安全性は，全レコードが完全に k 個ずつにクラスタ化され，一様な確率で推定する時の識別される顧客数の期待値で評価する．例えば，$k = 4$ の時，全体の 1/3 の顧客が再識別される．

表 4　RMSE 及び有用性結果

	有用性 (R)	有用性 (F)	有用性 (M)	有用性 (RFM)	安全性
2	0.270	0.463	0.352	0.097	0.50
3	0.214	0.343	0.301	0.040	0.33
4	0.155	0.287	0.288	0.026	0.25

4.2　考察

k の値が大きくなるほど有用性が下がり，安全性が上がった．R，F，M それぞれの有用性は R が最も低かった．その理由として，平均 117 日（$k = 2$）という購買日の区間の大きさが挙げられる．本研究は，より有用性を上げるために区間の開始日と終了日に元データの購買日を設定している．しかし，匿名加工データは区間の任意の購買日で評価している．そのため，元データとの誤差が大きくなったと考える．

RFM の有用性はどの k の場合も 1 割以下であった．しかし，R，F，M それぞれの有用性の積よりも高い有用性であったため，R，F，M は独立でないと考える．

5　おわりに

本研究では，購買履歴データについて，ユースケースを想定し，一般化の手法を用いた匿名加工データに対して有用性評価を行った．加工することにより，R，F，M それぞれの有用性は約 3 割，RFM の有用性は 1 割以下に減少した．本研究では区間に一般化されたデータの有用性を区間の任意の要素を選び評価した．そのため，評価するごとに有用性が異なるという問題点がある．今後は匿名加工データの RFM 計算を複数回行う等の対策を行い，より厳密な有用性の評価を行うことを課題とする．

参考文献

[1] 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.

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