一般化匿名加工された購買履歴データの顧客・商品の RFM 分析

小林祐貴 †

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

2017 年 5 月に個人情報保護法が改正され，中小企業をはじめとする全ての事業者が個人情報保護法の対象となった．また，一定の条件の下で加工を行うことにより，本人の同意がなくても第三者に提供・目的外利用を行うことができる匿名加工情報が新設された．匿名加工は，データから個人を特定されないように個人情報に対して加工を行うことである．加工処理において，匿名加工データから個人を特定されるのを困難にさせ，安全性を高めることが重要である．しかしながら，加工をし過ぎてしまうと，有用なデータからは遠のいてしまう．

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．一般化は，加工対象になる情報に含まれる記述について，上位の概念に置き換えること，又は数値を区間に置き換えることである．本コンテストでは参加者から集めた加工データを元データと比較し，平均誤差による一般的な有用性評価を行う．従って，加工データの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[3] の Online Retail DataSet(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，541909 レコード) のデータのうち，PWSCUP2018 で用いられた 81776 レコード 5 属性のデータを用いる．表 1 に本研究で使用する属性を示す．本稿では，これらを顧客 ID，時刻を削除した購買日，商品 ID，単価，購買数量と呼ぶ．表 2 に購買履歴履歴データの例を示す．

表 1　本研究で使用する属性

属性名	本稿での呼称
CustomerID	顧客 ID
InvoiceDate	購買日
StockCode	商品 ID
UnitPrice	単価
Quantity	購買数量

表 2　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
17042	2011/4/6	22439	0.65	10
15039	2011/5/9	21974	1.45	3
14911	2011/9/30	22818	0.42	12

2.2　オンライン購買履歴データの分析

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 4 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 3　購買総額上位 5 位の顧客 ID と購買金額

顧客 ID	12415	13694	12931	16422	12748
購買総額	124919	61908	42055	34684	32649

†Yuki Kobayashi, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

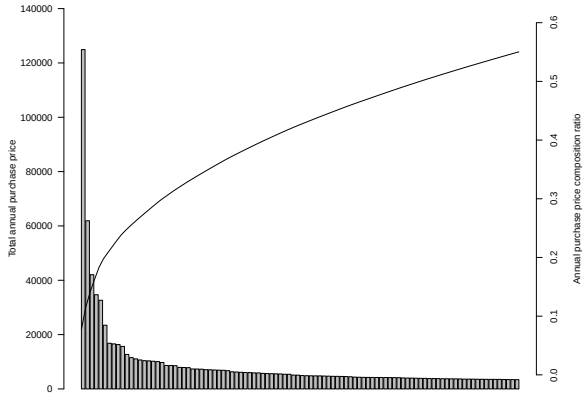

図 1　顧客ごとの年間購買総額と累積購買金額構成比率

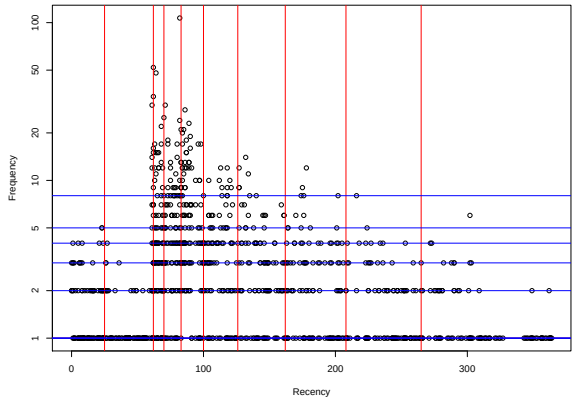

図 2　顧客の最新購買日と年間購買頻度を表す散布図

表 4　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

3　購買履歴データの匿名加工

本研究では PWSCUP2018 のルールに従い，k-匿名化を行う．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．PWSCUP2018 では，以下の加工が許されている．

(1) 維持（データの要素をそのまま利用）
(2) 削除（データの要素を削除し，削除したことを示す値に＊置き換える）
(3) 顧客 ID の仮名化（顧客 ID を任意の値へ変更すること）
(4) 一般化（与えられたデータの要素を区間や集合へ変更すること）．一般化の加工については以下のように属性値によって異なる．

（a）商品 ID の一般化商品 ID のように，属性値カテゴリ値である場合は，その顧客が年間に購買した商品の集合から，その属性値を含む部分集合へ一般化を行うことができる．以下に顧客 ID12348 が購入した商品 ID22423 を一般化する例を示す．顧客 ID12348 が年間に購買した商品の集合を $D = \{10002, 10120, 10125, 22423\}$ とする

$$22423 \rightarrow \{22423, 10120, 10125\}$$

（b）購買日，単価，購買数量の一般化属性値が購買日のように値の差には意味があるが比には意味がないとき．または，単価や数量のように比にも意味があるときは，元の値を含む単一の閉区間への一般化を行うことができる．以下に例を示す．

$$2011/1/30 \rightarrow [2010/12/15, 2011/2/10]$$

$$10.5 \rightarrow [5.0, 12.0]$$

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，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 5 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 5　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 購買履歴データの RFM と有用性評価・安全性評価

4.1 RFM の計算

本研究において，匿名加工データから RFM の計算を行う際，区間化，または集合に一般化されたデータから任意の値を選ばなければならない．そのため，本研究では以下のように一般化されたデータから値を選出する．

4.1.1 商品 ID の場合

名義尺度である商品 ID が一般化により集合になっている場合は，集合の中から任意の値 d を選出する．例えば，

$$d = 22500 \in \{22969, 22500, 22197\}$$

として評価する．

4.1.2 購買日の場合

間隔尺度である日付が一般化により区間化されている場合は，区間化された日付から一様に任意の日付を選び出し，区間開始日との日付差を計算する．これを n 回繰り返し，日付差の平均値を足しあわせた日付を区間化された購買日 $\overline{d}$ とする．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．例えば，$n = 3$ で $d_1 = 2011/1/1$，$d_2 = 2011/1/3$，$d_3 = 2011/1/2 \in [2011/1/1, 2011/1/3]$ の時，

$$\overline{d} = \frac{1}{3}\sum_{i=1}^{3} d_i = 2011/1/2$$

である．

4.1.3 単価，購買数量の場合

比例尺度である単価，購買数量が一般化により区間化されている場合は，区間化されている値から任意の値を選び，この操作を n 回繰り返し，平均値を区間化された比例尺度の値とする．例えば，$n = 3$ で $d_1 = 3, d_2 = 5, d_3 = 1 \in [1.0; 5.0]$ の時，

$$\overline{d} = \frac{1}{3}\sum_{i=1}^{3} d_i = 3.0$$

である．

4.2 有用性評価システム

図 3 に有用性評価システムの構成図を示す．$n = 100$ で実験した．

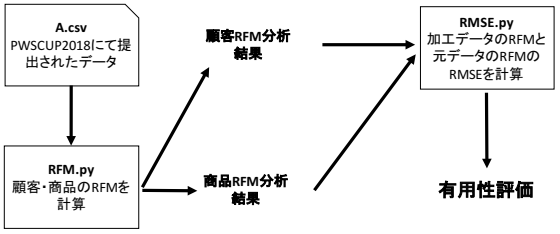

図 3　有用性評価システム構成図

4.3 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．

表 6 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．

4.4 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

表 6　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

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] 菊池浩明，他，“PWSCUP:履 歴データを安全に加工せよ”，コンピュータセキュリティシンポジウム (CSS2016)，pp.271-278，2016.

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