明治大学総合数理学部

2018 年度

卒　業　研　究

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

学位請求者　先端メディアサイエンス学科

小林 祐貴

目次

第 1 章

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[2] 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，PWSCUP2018 にて提出された匿名加工データの有用性評価を行う．

第 2 章

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[?] 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 2.1 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 2.2 に購買履歴履歴データの例を示す．

表 2.1　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 2.2　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 2.5 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 2.1 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．図 2.2 に商品ごとの年間売上総

額と累積売上構成比率を示す. 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．商品の売上分析では, 約 2 割の商品が全体売上の約 8 割の構成を占めており, パレートの法則が成り立っていた. パレートの法則は, 全体の要素のうち一部の要素が全体の大部分を占めている状態のことである. 表 2.3 に購買金額上位 5 位の顧客 ID と購買金額を示す. 顧客 ID12415 の顧客は年間に 124914 ポンド購買を行っている. 表**??**に年間売上総額上位 5 位の商品 ID と売上金額を示す. 商品 ID22423 の商品は年間に 124914 ポンドの売上を上げている.

図 2.3 に顧客の最新購買日と年間購買頻度の分布を示す．図 2.4 に商品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．本データでは, 購買履歴データの最新日付 (2011/11/30) から約 60 日前 (2011/9/30) に多くの常連客が購 買を止めていることがわかる. しかしながら, オンライン小売店は法人顧客であるため, 購買をやめたから離反客になったとは言えない.

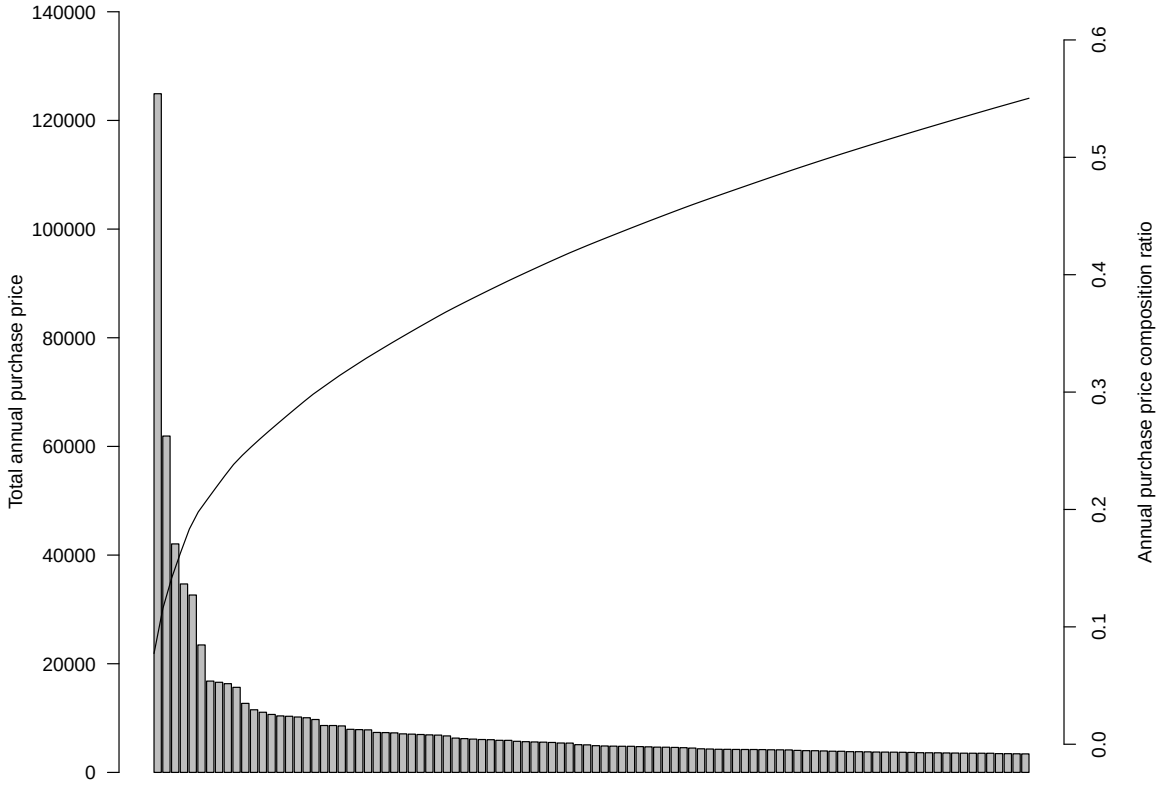

図 2.1　顧客ごとの年間購買総額と累積購買金額構成比率

表 2.3　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

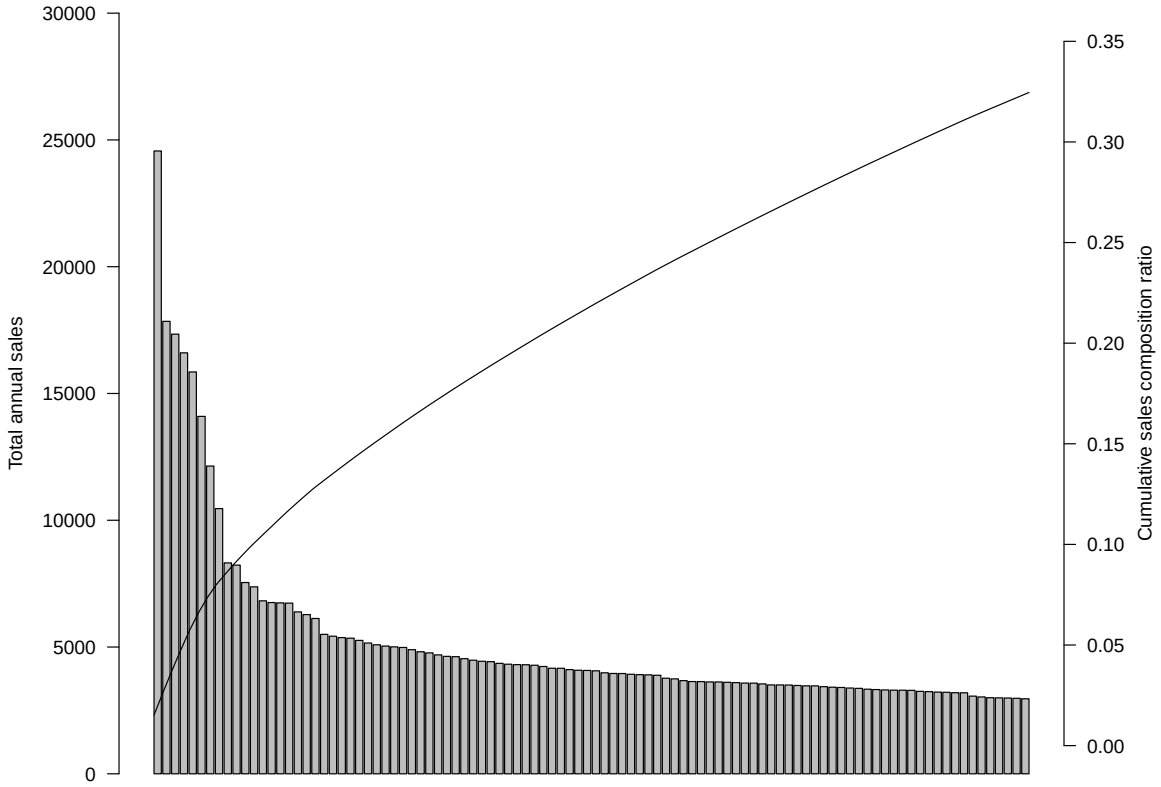

図 2.2　商品ごとの年間購買総額と累積購買金額構成比率

表 2.4　売上総額上位 5 位の商品 ID と購買金額

商品 ID	22423	85099B	84879	85123A	23084
売上総額	23564	17845	17340	16603	15849

表 2.5　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

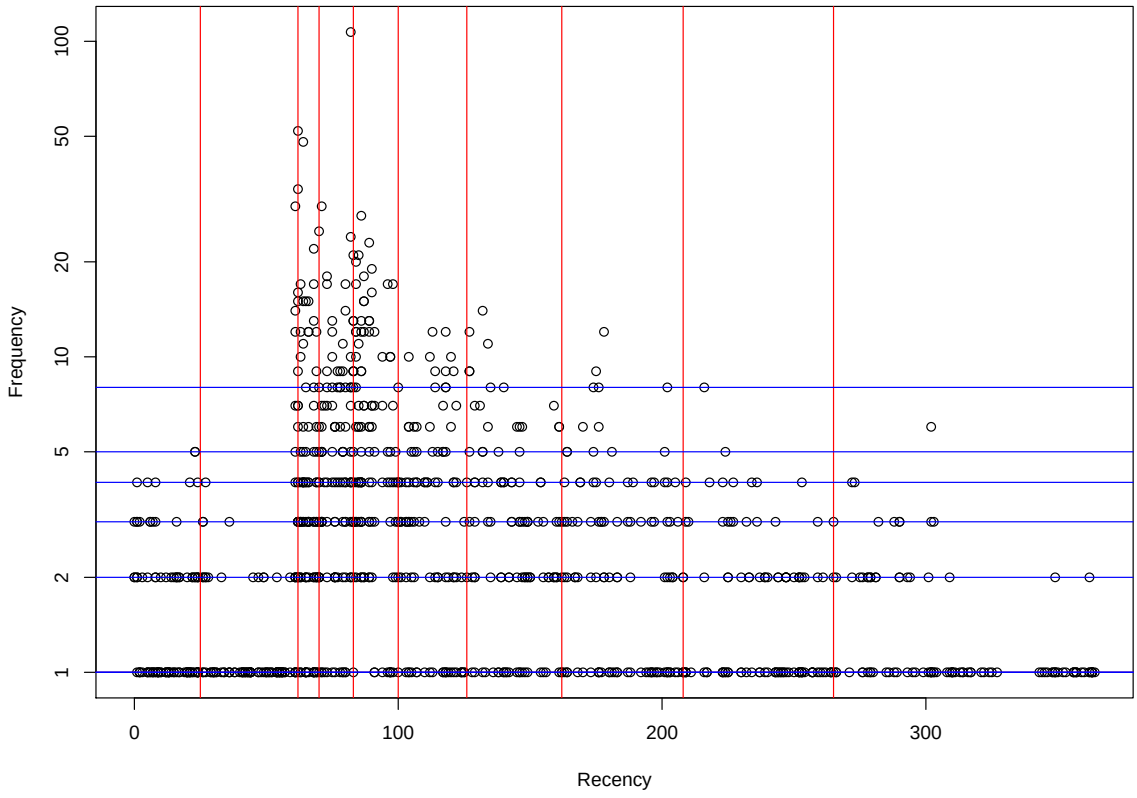

図 2.3　顧客の最新購買日と年間購買頻度を表す散布図

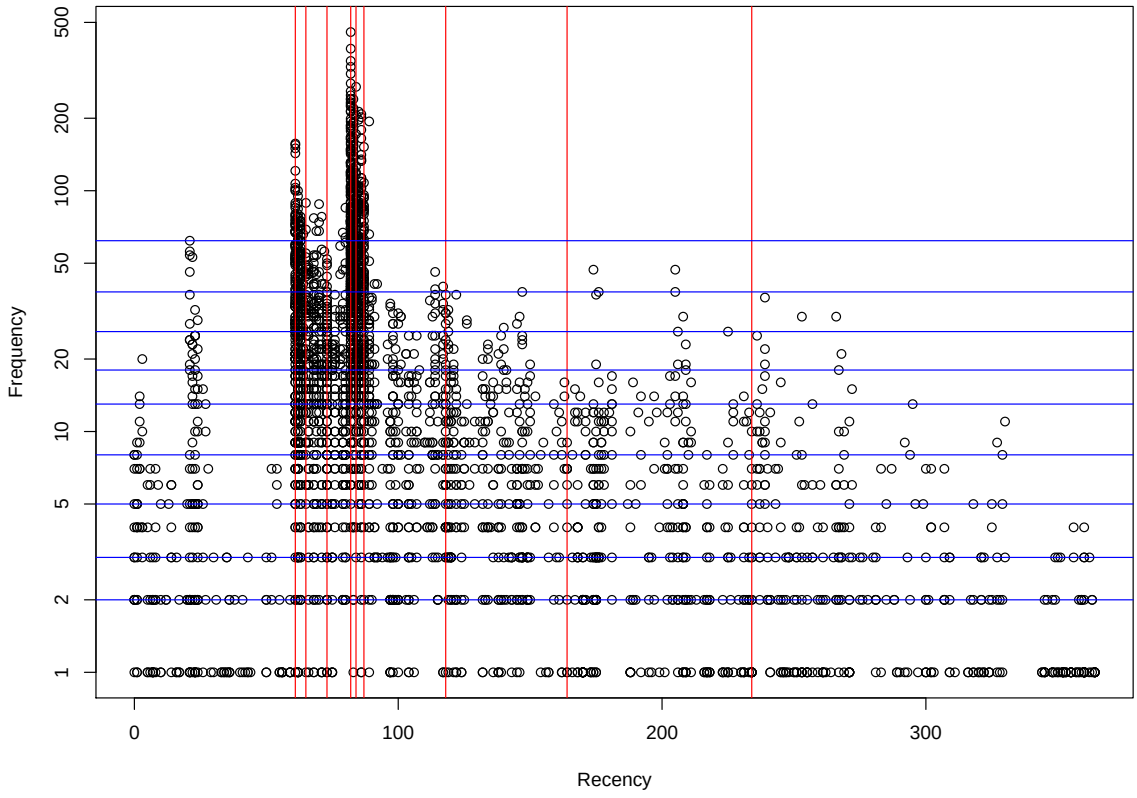

図 2.4　商品の最新購買日と年間購買頻度を表す散布図

第 3 章

購買履歴データの匿名加工

表 3.1　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]

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）

$$2011/1/1 \rightarrow 2011/1/1$$

(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）

$$2011/1/1 \rightarrow *$$

(3) 顧客 ID の仮名化（顧客 ID を任意の値へ変更すること）

$$12428 \rightarrow 23$$

(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 A.1 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．

第 4 章

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

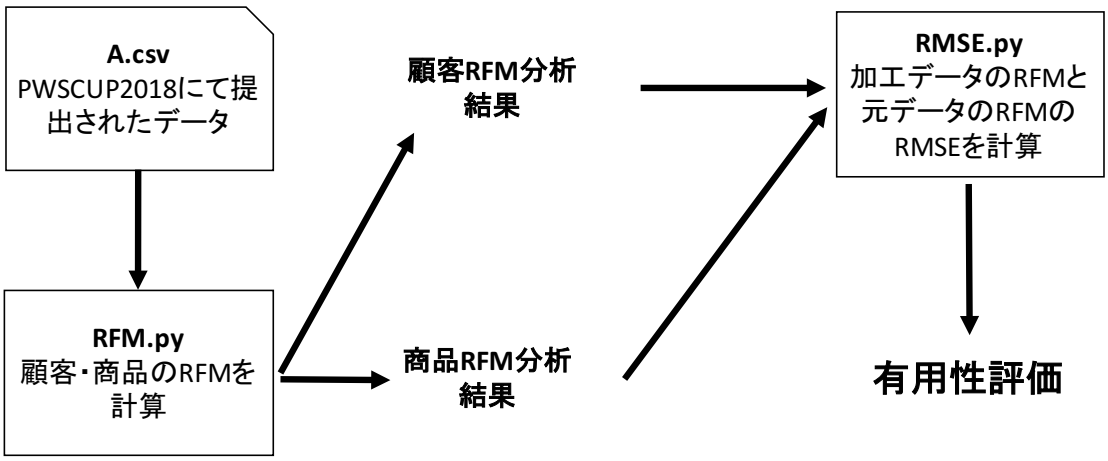

図 4.1　有用性評価システム構成図

4.3　有用性評価・安全性評価

表 4.1　有用性結果・安全性結果

	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

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 4.1 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．$k = 2$ の時は，R，F，M 一次元の有用性がそれぞれ 0.270，0.436，0.352，RFM 三次元の有用性は 0.097，安全性は 0.50 である．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，$k = 3$ 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．

第 5 章

PWSCUP2018 匿名加工データの有用性評価

5.1　有用性評価

PWSCUP2018 に提出された匿名加工データは，2 章の分析結果から以下のようにユースケースを想定し，それらを総合して匿名加工データの有用性評価を行う．

(1) 顧客ごとの RFM 分析による顧客判別購買履歴から，顧客ごとの最新購買日 (Recency)，購買頻度 (Frequency)，年間購買総額 (Monetary) の 3 つの指標を計算し，分析を行うことにより，新規顧客や常連顧客，離反顧客の判断を行うことができる．これにより，顧客に対して最適なマーケティング戦略を打ち出すことができる．

(2) 商品ごとの RFM 分析による商品判別購買履歴から，顧客ごとの最新購買日 (Recency)，購買頻度 (Frequency)，年間売上総額 (Monetary) の 3 つの指標を計算し，商品のカテゴライズを行うことで，商品発注や販売の際に，人気商品や不評商品の管理を行いやすくなる．

RFM はそれぞれ 10 分位値を目安に 10 ランクに分け，計 1000 ランクに顧客をクラスタリングし，元データと匿名加工データの二乗平均平方根誤差 RMSE で有用性評価を行う．ただし，有用性評価は RMSE の最大値を 1，最小値を 0，として [0，1] に正規化した値を用いる．以下に f を元データの RFM クラスタリング値，y を匿名加工データのクラスタリング値として，二乗平均平方根誤差 $RMSE$ の定義式を示す．

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2}$$

5.2　考察

表 5.1　RMSE 及び有用性結果

チーム	1	2	3	4	5	6	7	8	9	10	11	12	13	14
RMSE(顧客 RFM)	0.026	0.380	0.711	-	1	0.023	0.220	0	0.078	0.035	0.046	0.254	0.011	0.038
RMSE(商品 RFM)	0.320	0.778	0.248	-	0.452	0.329	0.350	0.360	0.380	0	0.335	1	0.158	0.338
有用性	0.346	1.158	0.959	-	1.452	0.352	0.570	0.360	0.458	**0.035**	0.381	1.254	0.169	0.376
PWSCUP2018 有用性	0.206	0.365	0.294	-	0.262	0.209	0.457	0.277	0.430	**0.193**	0.206	0.476	0.255	0.206

表 5.1 に有用性結果を示す．まず，PWSCUP2018 では，どのチームも k-匿名化を行っていた．k-匿名化

は，データに対して同一属性を持つデータを k 件以上になるようにデータを変更することで個人が特定される確率を k 分の 1 以下に低減することである．

本結果ではチーム 10 の有用性が最も高い．チーム 10 の有用性が高くなった要因は 3 点あると考えられる．1 点目は，2-匿名化を行っていたことだ．2-匿名化を行うことにより，1/2 の確率で顧客が識別されてしまう．しかしながら，顧客 2 名のレコードを同一にするので，3-匿名化や 4-匿名化を行うよりも元データとの誤差は低くなる．PWSCUP2018 では，元データと加工データの誤差で有用性評価しており，誤差の最も低いチーム 10 の加工データは RFM クラスタリングの誤差も低くなると考えられる．本結果で有用性の高い上位 5 チームは 2-匿名化を行っている．2 点目は，最適な顧客ペアを作っていることだ．チーム 10 は，2-匿名化の顧客ペアを作る際に，全通りの顧客ペアの有用性を計算し、最も有用性の高い顧客ペアで 2-匿名化を行っている．チーム 1，チーム 11，チーム 14 も同様の手法で加工データを作成しており，本結果でも高い有用性であった．3 点目は，レコード数上位 2 顧客のレコード削除を行っていないことだ．多くのチームが全顧客を加工していたのに対し，チーム 10 では PWSCUP2018 での有用性を上げるためにレコード数上位 2 顧客のレコード削除を行っていない．従って，元データとの誤差が少なくなり，有用性が高くなった．以上から，

(1) 2-匿名化
(2) 最適な顧客ペアで匿名加工
(3) レコード数上位 2 顧客のレコード削除不採用

の 3 点によってチーム 10 の本結果の有用性が高くなったと考えられる．

第 6 章

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．PWSCUP2018 に提出された匿名加工データの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．

謝辞

本研究を行うにあたり，多くの方より御指導いただきました．特に，多大なる御指導を受け賜りました，明治大学総合数理学部先端メディアサイエンス学科，菊池浩明教授に深く感謝申し上げます．予備実験等に協力してくださった菊池研究室の皆様並びに先端メディアサイエンス学科の方々に深く感謝の意を表するとともに，謝辞とさせていただきます．

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.

付録 A

購買履歴データにおける商品のアソシエーション分析と匿名加工データの再識別手法の提案

A.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. こういった中でデータの匿名加工の重要性も更に高まってきた.

匿名加工処理において, 匿名加工データの再識別を困難にさせ, 安全性を高めることが重要である. 加工をし過ぎてしまうと有用なデータから遠のいてしまう. しかしながら, 有用性については加工データの使い道や使う企業によって異なるので統一した基準を定めるのは困難である. 安全性については匿名加工・再識別コンテスト PWSCUP2016[1] にて Jaccard 再識別 [2] が最も有効であることが示された. しかし,Jaccard 再識別だけでは安全性を測れないデータ項目が存在することが懸念される. 例えば, 単価や購入数などのデータ項目では Jaccard 再識別を使用できないため, 安全性を測ることができない.

そこで, 本研究では,PWSCUP2016 にて提出された匿名加工データについて,Jaccard 再識別とは違う方法で再識別を行い, 安全性の検証を行うことを目的とする.PWSCUP2016 で使われたデータを分析し, そこで得られた結果を用いて再識別手法を提案する.

A.2　オンライン購買履歴データの分析

A.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 の Online Retail Data Set (2010 年から 8 年 8 か月分の英国のオンライン小売店での購買履歴データ，8 属性，541,909 レコード) の顧客データ (Master400.csv) と購入履歴データ (Transaction-Customer400.csv) を利用する.

A.2.2　オンライン購買履歴データの分析

匿名加工データのどの部分で再識別するかを検討するために, 様々な観点からデータ分析を行う. 本研究では, 購買履歴データの「商品 ID」と「伝票 ID」に注目して分析を行う.

図 A.1 に購買履歴データにおける商品ごとの出現回数を示す.POST とは postage（送料）のことでり, 商品とは関係ない. 最も出現回数の高かった ID22423 の商品は“Regency Cakestand 3 Tier”であり, 購入履歴データの全伝票のうち 2 割ほどを占めていた. イギリスのインテリア関連の雑誌には必ず紹介されると言っていいほどの人気商品である. また,2 4 位の商品は“Round Snack Boxes Set Of 4 Woodland”, “Spaceboy Lunch Box”, “Dolly Girl Lunch Box”である.

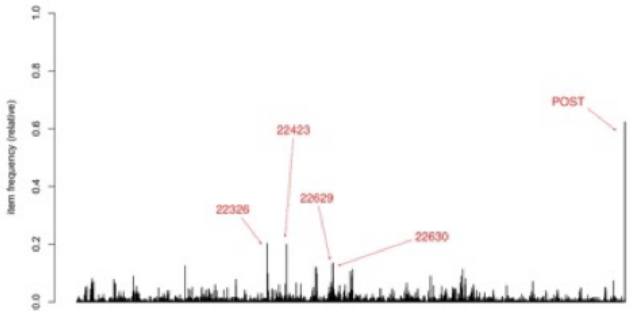

図 A.1　購買履歴データにおける商品ごとの出現回数

A.2.3　商品の連関規則分析

伝票 ID と商品 ID を用いて商品間の連関規則を調べる. 本研究では, バスケット分析を用いる. バスケット分析は,「客が買い物かごに一緒に入れる商品は何かを分析する」という例の様に, よく一緒に買われる商品を見つけるためのデータ分析手法である. 連関規則によって得られる商品間の関わりによって匿名加工データの再識別を行う.

結果の一部を表 4 に示す.lhs をルール・ヘッド (rule head),rhs をルール・ボディ (rule body) といい, それぞれ A,B で表す.support は全データの中で,「商品 A を買うときに, 商品 B も一緒に買う」（A ⇒ B）というルールが出現する割合を示しており, この結果はこの分析において不要となるデータを取り除く際の目安として用いる.confidence は条件部（A）の項目が出現する割合の中で,A と B が同時に出現する割合を示しており, この結果の値が高いほど「商品 A を購入する際に商品 B も一緒に買う」割合が高い. 表 4 を用いて説明すると,「商品 ID20750 を購入する際に商品 ID22326 も一緒に買う」確率が 0.64 だということを示している.lift

は商品 A と一緒に商品 B も購入した人の割合の, 全てのデータの中で商品 B だけを購入した人の割合に対する倍率である. 表 4 の例では，商品 ID20750 と一緒に商品 ID22326 も購入した人の割合は, 全てのデータの中で商品 ID22326 だけを購入した人より 3.14 倍の確率だったことを示している.

これらの連環規則より商品間の関わりを示した. しかしながら, まだ信憑性が低い. そこで, 本研究では confidence の値が高かった上位 100 点の商品を実際に調べる. ここで得られた特徴的な結果 2 点を図 A.2, 図 A.3 で示す. 図 A.2 は表 4 の規則 1 , 図 A.3 は表 4 の規則 2 である. 図 A.2 では「iPad の赤いケース A を買った人はお菓子の箱 B を同時購入する可能性が高い」ことを示している. 図 A.3 は,「赤い 1 0 個のライト A を買った人はケーキの乗せる皿 B を同時購入する可能性が高い」ことを示す. 同じ商品同士を同時購入しているルールが多い中, この 2 つの規則だけは特異な結果だと感じた.

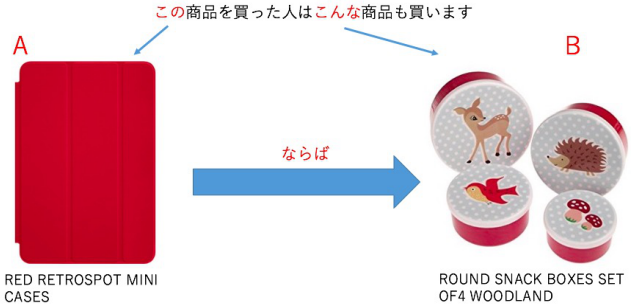

図 A.2　規則 1 の解釈

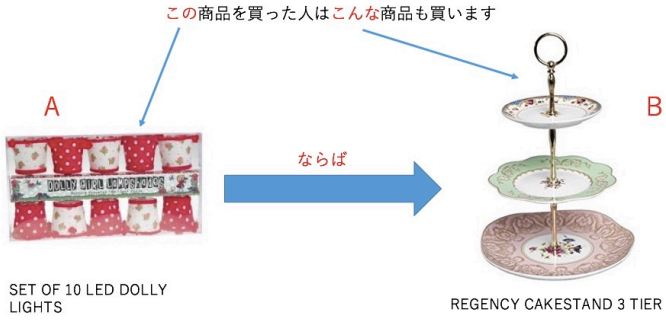

図 A.3　規則 2 の解釈

表 A.1　連関規則の結果の例

規則 No	lhs	rhs	support	confidence	lift
1	20750	22326	0.043	0.640	3.140
2	23108	22423	0.043	0.570	2.840

A.3　匿名加工データの再識別

A.3.1　匿名加工と再識別の概要

個人情報データの匿名加工についての概要を図４で示す．PWSCUP2016 において最も有効であった再識別手法である Jaccard 再識別法と比較する.Jaccard 再識別とは, 元データと加工データの商品 ID の集合の差分を求めて最も差の小さいデータで再識別する手法である.Jaccard は,

と定めている. ここで X,Y は元データと加工データの商品 ID の集合である.PWSCUP2016 での Jaccard 再識別の結果を表５第一列に示す.

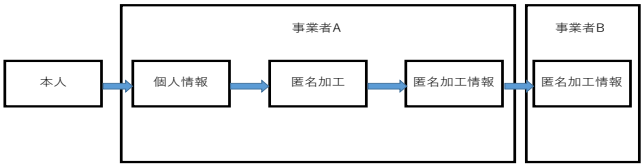

図 A.4　データ匿名加工の概要

A.3.2　Simpson 再識別

Jaccard 距離と同じく文章の類似度を調べる方法に Simpson 距離がある.Simpson 距離は相関関係と密接に関わっている. この係数が高ければ, 双方の集合の相関係数も強い.Simpson 距離の定義は以下の通りである.

集合 X と集合 Y が同一で, かつ一方が他方の部分集合である場合には 1 となる.

本研究では, この Simpson 距離を Jaccard 再識別に置き換えて「商品 ID」の集合に用いて再識別を行った. 結果を表５に示す. このデータで Jaccard 再識別には劣っている.Simpson 距離だと一方の要素数が極端に少ない場合でも高い算出結果を示すからだと考えられる. 例えば, 国語辞典という一個の文書を集合 X, 一語しか含まれていない文書を集合 X とした時, この係数は 1 に近づいてしまう.

A.3.3　ユークリッド距離再識別

本研究では次に単価と購入数量に注目する. 提案するユークリッド再識別では元データのベクトルと匿名加工データベクトル b の距離により再識別を試みる. ユークリッド距離を以下に示す.

本研究では,「単価」と「購入数量」の２つの項目を掛け合わせた「顧客一人当たりの平均購入額」と「顧客一人当たりの最大購入額」から成るベクトルのユークリッド距離から再識別を行う.

結果を表５に示す. ユークリッド再識は Simpson 再識別よりも再識別率が悪い. 原因は, 匿名加工データにて「単価」と「購入数量」の値を大きく変更されていたからだと考えられる. しかしながら, 対策をしていない加工データでは再識別率が高く,Simpson 再識別等と組み合わせればいい結果が得られるのではないかと考える.

A.3.4　コサイン類似度再識別

本研究では, ユークリッド再識別に代わってコサイン類似度を用いて 2 つのベクトルの角度の近さを計算し, 再識別を行う. コサイン類似度 cos(a,b) は,

で与えられる. ユークリッド距離と同様に顧客一人当たりの平均購入額と顧客一人当たりの最大購入額のコサイン類似度を用いて再識別を行う.

結果を表 5 に示す. 結果は悪く, 平均で 7 %ほどしか再識別ができていない. 原因についても未だ解明しきれていない.

A.3.5　結果の散布図

本研究では, 再識別の結果に関してどの程度 Jaccard 距離と結果が離れているのかを検討する.

各距離の散布図を図 5, 図 6, 図 7 に各々示す.

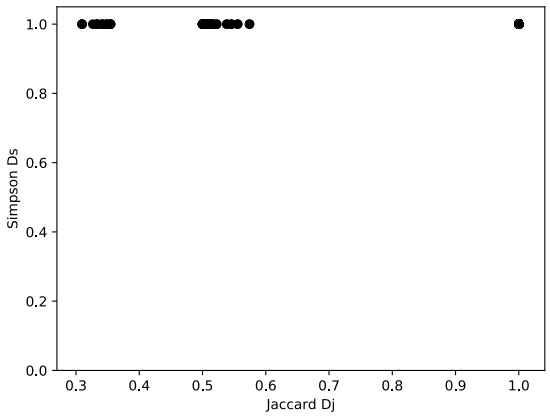

図 A.5　Jaccard 距離と Simpson 距離

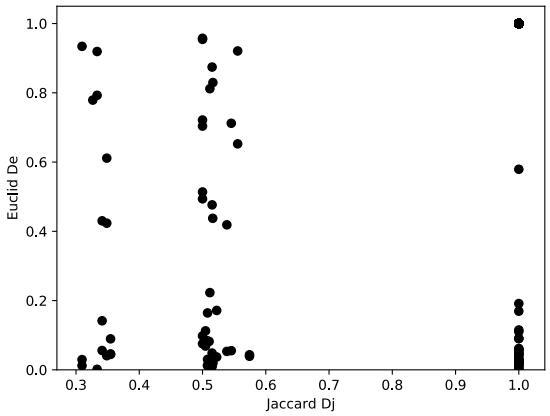

図 A.6　Jaccard 距離とユークリッド距離

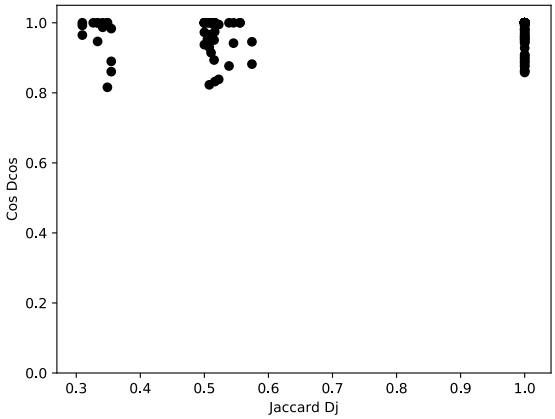

図 A.7　Jaccard 距離とコサイン類似度

A.4　おわりに

本研究では, 約４万レコードあるオンライン購買データ用いて分析を行った.PWSCUP2016 の匿名加工データを用いて再識別を行い, 加工データの安全性を確かめた. 本稿で提案した再識別手法はすべて Jaccard 再識別に勝ることができなかった. しかし, 提案した再識別手法を組み合わせることにより,Jaccard 再識別を超える再識別手法ができる可能性があると考えている. そのため, 今後の課題を連関規則にて得られた知見を用いた再識別手法と, 複数の再識別手法を組み合わせて新たな再識別手法を検討することとする.

参考文献

[1] 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.

[2] 原田玲央 “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”,pp.7-16,SCIS2017,2017.