乗降履歴データの有用性評価指標と匿名加工

伊藤聡志 †
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

企業は収集した顧客データやトランザクションデータを利活用する際，ユースケースに応じてリスク評価と匿名加工手法を考える必要がある．評価指標はデータの安全性や有用性を評価する指標であり，匿名加工は顧客データのような個人情報データから個人が特定されないように，データを加工することである．例えば，購買データを想定した評価指標・匿名加工に[1][2]がある．しかし，実際に顧客から同意を取りデータ収集するのは困難であった．

そこで，本研究の目的を実際に顧客から収集したデータを用いて匿名加工を行うこととする．31 人の交通 IC カードから乗降履歴データを取得し，そのデータのユースケースを想定して，それに対応する評価指標・匿名加工手法を検討する．

2　乗降履歴データの作成・分析

2.1　乗降履歴データの作成

本研究のために，明治大学総合数理学部に所属する 31 人の交通 IC カードから顧客データ M と乗降履歴データ T を作成した．なお，情報収集には Android のアプリケーション「IC カードリーダーby マネーフォワード[4]」を使用した．アプリケーションの仕様上，一人あたりから収集できる履歴は最大 19 件である．表 1 にアプリケーションで取得できる乗降履歴データ T の例を示す．

表 2 に取得した本データの概要を示す．顧客データ M(マスターデータ)は 31 レコード 6 属性のデータであり，乗降履歴データ T(トランザクションデータ)は 584 レコード 10 属性のデータである．表 3 に顧客データの例を示す．表 4 に乗降履歴データの例を示す．本来，交通 IC カードの利用履歴で得られる情報は「日付」,「利用内容」,「使用金額」の 3 属性のみであるが，本データでは「利用内容」属性を 6 属性に細分化している．例えば，表 1 の乗降履歴をデータ化したものが表 4 であるが，「利用内容」属性を「乗車駅」,「降車駅」,「乗車路線」,「降車路線」,「用途」,「使用場所」の 6 属性に分けている．「用途」属性には IC カードの用途(交通や物販等 5 種類)を示し，「使用場所」属性には IC カードを使用した場所(券売機や自販機等 8 種類)を示している．

顧客データ M は IC カードから作成できないため，顧客本人から情報を取得し作成した．定期券の区間で乗り降りした履歴は取得できないため，顧客データ M に定期券の範囲を加えた．

表 1　取得できる乗降履歴の例

日付	利用内容	使用金額
2016/10/30	入 上野 (JR 東北本線) 出 高田馬場 (JR 山手線)	-194
2016/10/30	入 高田馬場 (JR 山手線) 出 上野 (JR 東北本線)	-194
2016/10/8	チャージ 券売機等	2000

表 2　作成したデータの概要

	データ種別	データ件数	データ項目	項目
個人情報	顧客データ M	n 31 件	顧客 ID	2 桁数値
			性別	男女
			学年	1 桁数値
			住所	名称
			定期券範囲 1	名称
			定期券範囲 2	名称
	乗降履歴データ T	m 584 件	顧客 ID	2 桁数地
			日付	yyyy/mm/dd
			回数	数値
			乗車駅	名称
			降車駅	名称
			乗車路線	名称
			降車路線	名称
			用途	カテゴリ
			使用場所	カテゴリ
			料金	数値

表 3　顧客データ M の例

ID	性別	学年	住所	定期券範囲1	定期券範囲2
1	男	1	千葉県	NA	NA
2	女	3	東京都	中野	新宿

2.2　乗降履歴データの分析

3 節で示すユースケースへの適用可能性を明らかにするために，乗降履歴データの「使用料金」と「駅利用回数」に注目して分析を行う．

図 1 に月日ごとの使用料金の変化を示す．情報を収集したのが 6 月であることと，収集できる履歴が直近 19 件ま

†Satoshi Ito, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

表４　乗降履歴データTの例

顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
1	2016/10/30	2	上野	高田馬場	JR東北本線	JR山手線	交通	NA	-194
1	2016/10/30	1	高田馬場	上野	JR山手線	JR東北本線	交通	NA	-194
1	2016/10/8	1	NA	NA	NA	NA	チャージ	券売機	2000

でであることから，4〜6 月の使用料金が多くなっている．また，図 2 にユーザごとの総使用料金を示す．ユーザの総使用料金の統計量を表 5 に示す．

図 3 に上位 130 位の駅の利用回数を示す．被験者の所属する中野キャンパス周辺の中野駅や新宿駅の利用回数が非常に多く(4 位まで)，その他の駅の利用回数と大きな差が見られた．表 6 に利用回数上位 5 位の駅名と回数を示す．

図 5 にユーザ間の利用駅についての類似度の度合いを表す Jaccard 距離の分布を示す．本データの平均 Jaccard 距離は 0.933 であった．このことより，本データのユーザは利用駅について，ほとんど似ていないことがわかる．

表 7 に顧客属性(性別, 学年)のクロス集計表を示す．

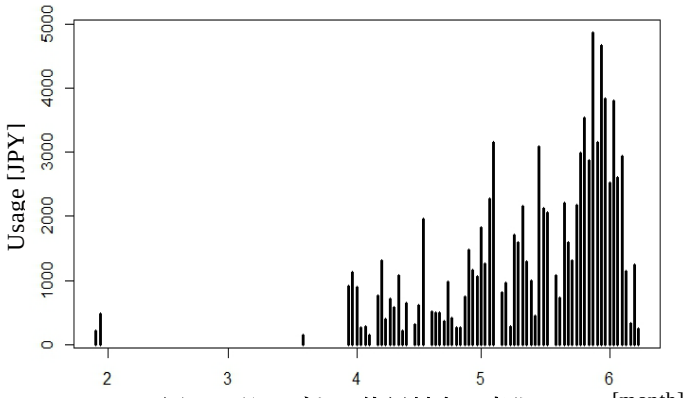

図１　月日ごとの使用料金の変化

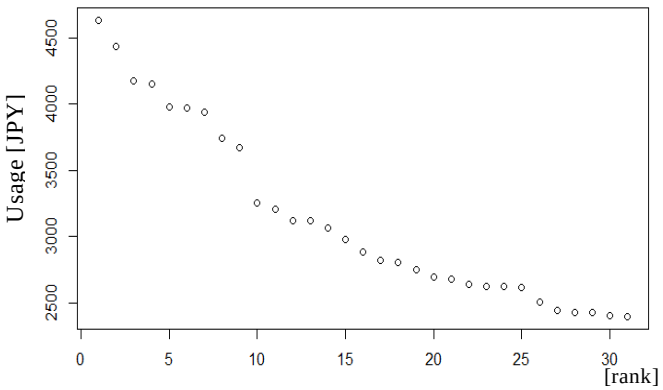

図２　全ユーザの総使用料金

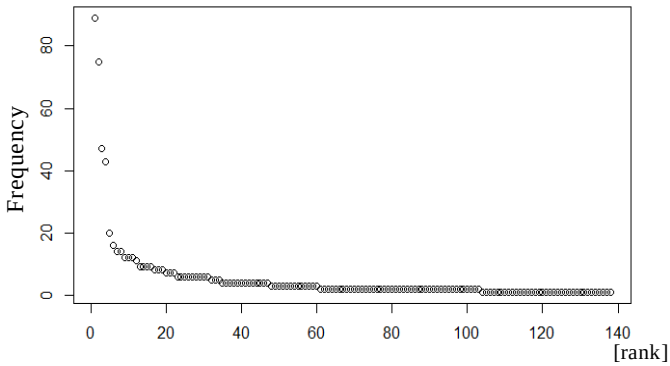

図３　駅の利用回数

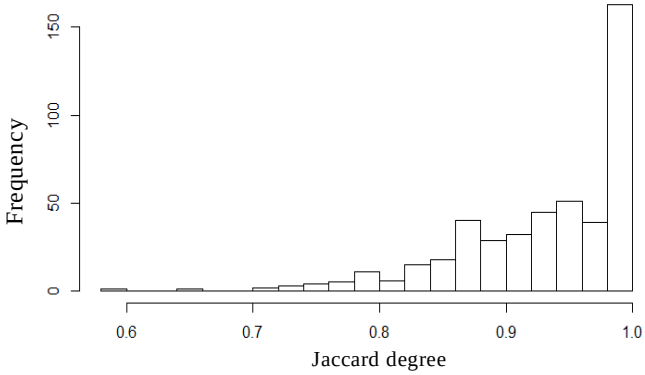

図５　ユーザ間の駅利用についての Jaccard 距離の分布

表５　総使用料金の統計量(円)

平均	3133.871
最大	4633
最小	2393

表６　利用回数上位の駅名と回数

新宿	中野	渋谷	高田馬場	明大前
89	75	47	43	20

表７　顧客属性のクロス集計表

性別\学年	0(教授)	1	2	3	4	計
男	1	6	5	4	10	26
女	0	2	0	2	1	5

3　乗降履歴データのユースケースと評価指標

本データのユースケースを表 8 のように想定する．本ユースケースは，本データを用いて明治大学総合数理学部に所属する人に対して広告・勧誘を行う効果的な場所を選定することを想定している．なお，ユースケースの作成には経済産業省の匿名加工情報作成マニュアル[3]を参考にした．例えば，外部組織が明治大学総合数理学部に所属する 3・4 年生の男性に対して広告・勧誘を行う場合，顧客属性が「性別=男，学年=3or4」の全ユーザの駅利用回数を用いる．

想定したユースケースに対応する評価指標を検討する．[2]の匿名加工データの有用性評価の多くは「元データの特徴をどれだけ保持しているか」という観点で評価されている．そこで，本ユースケースでは，以下の特徴を保持できているかどうかで匿名加工データの有用性を評価する．

1. 顧客属性(性別，学年)毎の駅利用回数(U1)
2. 駅利用回数の順位(上位のみ)(U2)
3. 顧客属性(性別，学年)のクロス集計の人数(U3)

また，これらの評価を行う有用性指標を順に U1, U2, U3 とする．各指標の式を以下に示す．M，T が元データを表し，M*，T*が加工されたデータを表す．$T_{station}(X_i)$は T についてのグループX_iの駅利用総回数を表す．g は T のグループ数を表す．S_5を上位 5 駅の集合とする．rank(T,s)は駅 s の T における利用回数順位を表している．$Cross_{sex,grade}$は M の(性別，学年)属性についてのクロス集計値を表し，$num(属性名)$はその属性の種類数を表す．これらの有用性指標の値が 0 に近いほど，データ(T*,M*)の有用性は高い．

$$U_1(M,T,M^*,T^*)=\frac{\sum_{g}^{i=1}|T_{station}(X_i)-T^*_{station}(X_i)|}{g}$$

$$U_2(M,T,M^*,T^*)=5-|\{s\in S_5(rank(T,s)=rank(T^*,s))\}|$$

$$U_3(M,T,M^*,T^*)=\frac{\sum_{num(sex)}^{i=1}\sum_{num(grade)}^{j=1}|Cross_{sex,grade}(i,j)-Cross^*_{sex,grade}(i,j)|}{num(sex)*num(grade)}$$

表８　想定するユースケース

匿名加工情報	顧客属性に応じた駅利用回数
業務サービス概要	明治大学総合数理学部に所属する人が利用している駅，またその回数を顧客属性(性別・学年)に応じて適した広告を配信する
提供する属性	M(顧客 ID，性別，学年) T(顧客 ID，乗車駅，降車駅，乗車路線，降車路線)
匿名加工情報利用目的	利用者に応じた最適な広告・勧誘を行うこと

4　乗降履歴データの匿名加工

顧客データや乗降履歴データをそのまま外部組織に提供してしまうと顧客個人が特定されてしまう場合がある．本ユースケースでは以下の場合が考えられる．

1. 顧客属性(性別，学年)の組み合わせが特殊である
2. 特殊な駅(利用回数が少ない，極端に離れた場所にある)を利用している

例えば本データの場合，表 7 より顧客属性が「性別=女，学年=4」であるユーザは 1 人しかいないため，個人が特定されてしまう．また，特殊な駅を利用している場合も個人が特定されやすい．例えば静岡駅を繰り返し利用している履歴があった場合，その履歴は住所が静岡県の顧客のものである可能性が高い．本節では，有用性を保ちつつ，これらの特殊な値を持たないデータを作成する手順を，簡易データを用いて説明する．表 9 に簡易顧客データ M，表 10 に簡易乗降履歴データ T を示す．この場合，3 つのグループ(A:性別=男，学年=1)，(B:性別=男，学年=2)，(C:性別=女，学年=4)ができる．

まず，前節で定義した有用性指標を損なわない加工手法を考える．U1 は駅利用回数についての有用性指標であるため，なるべく駅利用回数を保持する必要がある．しかし，顧客属性が同じグループ内で利用駅(乗車駅，降車駅)をシャッフルしても顧客属性ごとの駅利用回数は変化しないため，U1 を損なうことはない．全体の駅利用回数も変化しないため，U2 を損なうこともなく，顧客データは加工しないため U3 も損なわない．この手法を「グループ内シャッフル」とする．表 10 の乗降履歴データ T の利用駅をグループ内シャッフルした結果 T*を表 11 に示す．

次に個人が特定されないようにデータを加工する．グループ内シャッフルのみだと顧客データは無加工であるため，表 7 より「性別=男，学年=0」と「性別=女，学年=4」の顧客が容易に特定されてしまう．表 12 に加工した顧客データ M*を示す．この場合，顧客データの属性の組み合わせが特殊な顧客は(性別=女，学年=4)であるため，グループ B と同じ顧客属性(性別=男，学年=2)に加工する．

特殊な駅を利用している顧客も特定されやすい．これらを解決するために，顧客属性の組み合わせが特殊な顧客を別のグループに移し，特殊な駅の利用履歴を全て利用回数 1 位の「新宿」に置き換える．これらの加工では顧客属性ごとの駅利用回数や人数が変わってしまうため U1 と U3 を損なってしまうが，駅利用順位は変わらない(1 位の利用回数がさらに増えるだけ)ので，U2 は損なわない．表 13 に加工された乗降履歴データ T**を示す．この場合特殊な駅は「熱海」であるため，利用回数 1 位の「新宿」に置き換える．

以上の様にして，想定したユースケースに対応する評価指標を作成し，それを満たし，かつ個人が特定されにくい匿名加工データ M*，T**を作成した．表 14 に T, T*,

T**, M, M*についてのU1~U3の値の変化を示す．Mを加工したことによってU1, U3が上がってしまったが，U2の値は変化しておらず，有用性を保てている．

表９　簡易顧客データ M

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	女	4	C

表１０　簡易乗降履歴データT

顧客ID	乗車駅	降車駅	group
1	新宿	品川	A
1	品川	新宿	A
2	高田馬場	新宿	A
2	新宿	中野	A
3	中野	新宿	B
3	新宿	中野	B
4	高田馬場	品川	B
4	品川	熱海	B
5	中野	東京	C
5	東京	中野	C

表１１　グループ内シャッフルされた乗降履歴データT*

顧客ID	乗車駅	降車駅	group
1	**新宿***	**新宿***	A
1	**高田馬場***	**品川***	A
2	**品川***	**中野***	A
2	**新宿***	**新宿***	A
3	**品川***	**新宿***	B
3	**中野***	**品川***	B
4	**高田馬場***	**中野***	B
4	**新宿***	**熱海***	B
5	中野	東京	C
5	東京	中野	C

表１２　加工された顧客データM*

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	**男***	**2***	**B***

表１４　U1~U3の値の変化

	M,T	M,T*	M*,T*	M*,T**
U1	0	0	0	2.67
U2	0	0	0	0
U3	0	0	0.2	0.2

表１３　加工された乗降履歴データT**

顧客ID	乗車駅	降車駅	group
1	新宿	新宿	A
1	高田馬場	品川	A
2	品川	中野	A
2	新宿	新宿	A
3	品川	新宿	B
3	新宿	品川	B
4	高田馬場	中野	B
4	新宿	**新宿***	B
5	中野	東京	B
5	東京	中野	B

5　おわりに

31人の交通ICカードから取得した顧客データと乗降履歴データのユースケースを想定し，それに対応する評価指標と加工手法を考えた．本稿では説明のため，作成した乗降履歴データではなく簡易データに対して加工を行ったが，実際のデータにもこれらの手法は適用できる．本研究は非常に小規模なデータによるものであり，想定したユースケースも現実性に欠ける．

本稿では安全性をデータが特殊な値(顧客，利用駅)を持つ場合を想定し，それらを持つ場合は危険なデータ，持たない場合は安全なデータと定義している．しかし本来，データの安全性評価は様々な再識別手法を想定する必要があるため，非常に困難である．データの評価指標についての議論はPWSCUP[1][2]を通して行われている．

より大規模なデータの作成・分析や，それを用いたより具体的なユースケースの想定とそれに対応する評価指標や加工手法の提案・実装を今後の課題とする．

参考文献

[1] 菊池浩明，山口高康，濱田浩気，山岡裕司，小栗秀暢，佐久間 淳，“匿名加工・再識別コンテスト Ice & Fire の設計”, CSS 2015, pp.363-370, 2015.

[2] 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”, CSS2016, pp.271-278, 2016. (https://pwscup.personal-data.biz, 2016年12月参照.)

[3] 経済産業省，事業者が匿名加工情報の具体的な作成方法を検討するにあたっての参考資料(「匿名加工情報作成マニュアル」)Ver1.0 (http://www.meti.go.jp/press/2016/08/20160808002/20160808002-1.pdf, 2016年12月参照.)

[4] ICカードリーダーbyマネーフォワード (https://play.google.com/store/apps/details?id=com.moneyforward.nfcreader&hl=ja)