Vol.2020-DPS-182 No.4
Vol.2020-CSEC-88 No.4
2020/3/12

履歴データの数理モデルの提案とk-匿名化に必要なダミーレコード数推定への応用

伊藤 聡志[1,a)]　菊池 浩明[2,b)]

概要：企業などの組織が収集したビッグデータ (顧客情報や位置情報など) を安全に分析・利活用するためには，データの匿名加工とリスク評価が不可欠である．匿名加工は個人が識別されることを防ぐために個人情報を加工する技術であり，代表的なものとして k-匿名化がある．顧客ごとにレコード数が異なるデータ (履歴データ) を k-匿名化するためには，ダミーレコードを追加して複数顧客のレコード数を揃える必要がある．追加するダミーレコードの数が多くなるほどデータの有用性は下がるが，その数は実際に加工するまでわからないという問題がある．本稿では，履歴データの項目（購買商品や移動位置）が一様分布する仮定の下，履歴データ中に登場する項目の種類数の確率分布とその期待値を与える数理モデルを提案する．本モデルを用いることにより，元データの統計量 (レコード数や顧客数など) と加工パラメータ k といった加工前に知ることのできる値から，履歴データの k-匿名化に必要なダミーレコードの数が推定できる．

キーワード：匿名加工，履歴データモデル，k-匿名化，ダミーレコード

Mathematical model of Transaction Records and its Application to Estimate Additional Dummy Records for k-anonymity

1. はじめに

企業などの組織が収集したビッグデータ (購買情報や位置情報など) を安全に分析・利活用することにより，我々は様々な恩恵を得ることができるが，そのためにはデータの匿名加工とデータの評価が不可欠である．匿名加工は個人が識別されることを防ぐために個人情報を加工する技術であり，匿名加工されたデータから個人を識別しようとする攻撃を再識別という．匿名加工されたデータは有用性・安全性の 2 点で評価されることが多く，評価指標が匿名加工・再識別コンテスト PWSCup[1] 等で盛んに研究されている．

代表的な匿名加工手法として k-匿名化 [2] という手法があり，これはデータ中の少なくとも k 人が同じ情報を持つ (区別がつかない) ようにデータを加工する手法である．顧客ごとにレコード数が異なるデータ (履歴データ) を k-匿名化するためには，レコード削除や疑似レコード (ダミーレコード) の追加が必要になる．例えば 2 レコードを持つ顧客と 3 レコードを持つ顧客の区別をつかなくするためには，疑似レコードを 1 つ追加して 2 人とも 3 レコードを持つようにしてから一般化等の加工を行えばよい．我々の過去の研究 [4] では，顧客 400 人分 38,087 レコードの購買履歴データを Jaccard 距離を用いて 5-匿名化するためには，疑似レコードが約 170,000 レコード必要であるということが明らかになっている．また PWSCup 2018[3] では，用いられたデータのレコード数の多い上位 2 顧客の区別をつかなくするために，疑似レコードが 2,688 レコード必要であった．

このような「疑似レコード数」は，加工されたデータの有用性評価指標のひとつとして考えられる．例えば元データに 100 個の疑似レコードを追加したデータと 1,000 個の疑似レコードを追加したデータがあった場合，有用性が高いのは前者であるといえる．そのため，疑似レコード数は加工データにとって重要な値であるといえるが，この値はパラメータ k を決めて実際にデータを加工するまで求める

[1] 明治大学大学院先端数理科学研究科
Nakano, Nakano-ku, Tokyo 164–8525, Japan
[2] 明治大学総合数理学部
Nakano, Nakano-ku, Tokyo 164–8525, Japan
[a)] mmhm@meiji.ac.jp
[b)] kikn@meiji.ac.jp

ことはできない．しかし，最適な k の値を求めるためには，有用性指標として様々な k で加工をした際の疑似レコード数を求める必要がある．

そこで，本研究の目的を，履歴データを加工する前に k-匿名化に加工が必要な疑似レコード数を推定することとする．この問題は，次のように定められる．(問題)「ℓ 種類の値をとる x レコードの履歴データが持つ値の種類数 y」を求めよ．ここで，履歴データ中のある属性の値が全て一様確率で生起すること ($1/\ell$ の一様分布)，履歴データ中の全ての顧客のレコード数が等しいこと，履歴データ中の全てのクラスタの大きさが等しいことの 3 つを仮定する．また，本稿では Y の期待値を考える．

本研究と似た議論として，クーポンコレクター問題 [6] が知られている．(クーポンコレクター問題) 各箱には ℓ 種の異なるクーポンが入っている．n 種のクーポンの選択は独立で，一様に分布しているとき，全種類を少なくとも 1 種以上集めるには何箱買えばよいか？この問題は，X_i を i 番目のクーポンを集めるのに必要な箱の数を表す確率変数と置くと，$X = \sum_{i=1}^{\ell} X_i$ を求める問題であると考えることができる．X_i は確率 $p_i = 1 - \frac{i-1}{n}$ の幾何分布に従うので，その期待値は $E[X_i] = 1/p_i = \frac{\ell}{n-i+1}$ である．期待値の線形性より，$E[X] = \sum_{i=1}^{\ell} E[X_i] = \sum_{i=1}^{\ell} \frac{\ell}{\ell-i+1} = n\sum_{i=1}^{\ell} \frac{1}{i}$ であり，ハーモニック数 $H(n) = \sum \frac{1}{i} = \ln \ell + \mathcal{O}(1)$ から，クーポンコレクター問題は $E[X] = \ell \ln \ell + \mathcal{O}(\ell)$ と解ける．

クーポンコレクター問題と本問題の違いを整理し，表 1 に示す．クーポンコレクター問題と本問題のゴールは異なるため，これを本研究に適用することはできない．そのため本研究では，履歴データ中の値の種類数の確率分布を与える数理モデルを提案する．このモデルを用いることによって，x レコードの履歴データが持つ値 (全 ℓ 種類) の種類数 y の期待値を求めることができ，提案モデルを応用することによって，加工前に手に入る値 (元データの統計量やパラメータ k) から疑似レコード数の期待値を求めることができる．

本稿では，2 章で履歴データモデルの提案と分析を行い，3 章で提案モデルを用いて k-匿名化に必要な疑似レコード数の期待値を求める．

2. 履歴データモデル

2.1 基礎定義

本研究では，レコード（行）と属性（列）によって構成される履歴データを考える．記号等を以下のように定義する．

定義 2.1 データを T とする．T はレコード数 m，顧客数 n のデータであり，顧客 ID 属性と ℓ 種類の値をとる属性からなるものとする．T の顧客集合を $U = \{u_1, \ldots, u_n\}$ とし，U が属性でとる値の集合を $I(U) = \{g_1, \ldots, g_\ell\}$ とする．

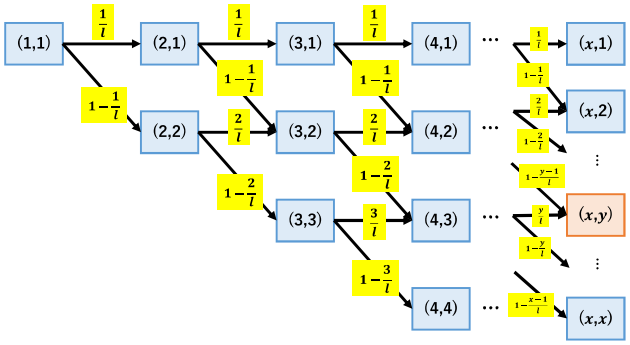

図 1 (x, y) の状態遷移図

例 2.1 T の例として，3 人の顧客の購買商品のデータ T_{ex} を表 2 に示す．この場合，$n = 3, m = 4, \ell = 2, U = \{Alice, Bob, Carol\}$ である．T_{ex} 内では Apple,Book が購買されているので，$I(U) = \{Apple, Book\}$ であり，Alice の購買商品集合は $I(Alice) = \{Apple\}$ である．

定義 2.2 ℓ 種類の属性の値は $1/\ell$ の独立で一様な確率で生起する ($1/\ell$ 仮定)．

定義 2.3 ℓ 種類の値をとる属性のデータが x レコードあるときに，値が y 種類ある状態を (x, y) とし，その状態になる条件付確率を $Pr(Y = y|X = x)$ とする．また，履歴データが x レコードあるときの属性 (全 ℓ 種類) の種類数の期待値を $f(x, \ell)$ とする．

(x, y) になるパターンは，$(x-1, y)$ から確率 y/ℓ で値が重複する場合と，$(x-1, y-1)$ から確率 $1 - \frac{y-1}{\ell}$ で新しい値が生起する場合の 2 パターンあるため，$Pr(y|x)$ は $Pr(y-1|x-1), Pr(y|x-1)$ を用いて漸化式

$$Pr(y|x) = (1-\frac{y-1}{\ell})Pr(y-1|x-1)+\frac{y}{\ell}Pr(y|x-1)(1)$$

と表せる．

例 2.2 (x, y) の状態遷移図を図 1 に示す．例えば購買履歴データの場合，$(4, 2)$ はデータが 4 レコードあるときに，商品が全 ℓ 種類中 2 種類生起している状態である．図 1 より，$Pr(Y = 2|X = 4) = ((1/\ell)(1/\ell)(1-1/\ell)) \cdot ((1/\ell)(1-1/\ell)(2/\ell)) \cdot ((1-1/\ell)(2/\ell)(2/\ell)) = 7(1-1/\ell)/\ell^2$ であり，$f(4, \ell) = \sum_{i=1}^{4} i \cdot Pr(i|4)$ である．

2.2 履歴データモデルの提案

本節では，$1/\ell$ 仮定のもと，履歴データ中に登場する項目の種類数の確率分布 $Pr(y|x)$ とその期待値 $f(x, l)$ を与える数理モデルを提案する．

$Pr(y|x)$ は以下のように求めることができる．

定理 2.1 ℓ 種類の値をとる x レコードのデータの属性中に，y 種類の値がある生じる確率 $Pr(y|x)$ は

$$Pr(y|x) = \prod_{j=0}^{y-1}(1-\frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y} (2)$$

である．ただし，$m_1 \ldots, m_y \geq 0, x \geq y \geq 1$ である．

表 1　クーポンコレクター問題と本問題との違い

	クーポンコレクター問題	本問題
値の生起	$1/\ell$ 一様性	$1/\ell$ 一様性
ゴール	全 ℓ 種を得るための箱の数 X	$x < \ell$ 個のレコードがあるときの値の種類数 Y
期待値	$\ell \ln \ell + \mathcal{O}(\ell)$	?

表 2　履歴データ T の例 T_{ex}

顧客 ID	購買商品
Alice	Apple
Bob	Apple
Bob	Book
Carol	Book

(証明)　式 2 を数学的帰納法によって証明する．

$x = 1$ のとき，$x \geq y \geq 1$ より $y = 1$ であり，$Pr(1|1) = (1 - \frac{0}{\ell}) \cdot (\frac{1}{\ell})^0 \cdots (\frac{y}{\ell})^0 = 1$ となるため，式 2 が成り立つ．

$x = x' - 1 (x' \geq 2)$ のとき，任意の $y (1 \leq y \leq x' - 1)$ で式 2 が成り立つと仮定する．つまり，$Pr(y|x' - 1) = \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=(x'-1)-y}(\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y}$ とする．このとき，$Pr(y-1|x'-1) = \prod_{j=0}^{y-2}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_{y-1}=(x'-1)-(y-1)}(\frac{1}{\ell})^{m_1} \cdots (\frac{y-1}{\ell})^{m_{y-1}}$ となる．

これらを $x = x'$ のときの式 1 に代入すると，式 2 を得る．

$$\begin{aligned}
Pr(y|x') &= (1 - \frac{y-1}{\ell})Pr(y-1|x'-1) + (\frac{y}{\ell})Pr(y|x'-1) \\
&= \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_{y-1}=x'-y}(\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_{y-1}} \\
&\quad + \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x'-y}(\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y+1} \\
&= \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x'-y}(\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y}
\end{aligned}$$

よって，$x = x'$ のときも式 2 は成り立つため，任意の $x (x \geq 1)$ について式 2 は成り立つ．　(Q.E.D)

補助定理 2.1　x レコードの履歴データが持つ値の種類数 (全 ℓ 種類) の期待値は $f(x, \ell) = \sum_{i=1}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i-1}$ である．

補助定理 2.2　$\sum_{i=1}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i-1} = (-\ell)(1 - \frac{1}{\ell})^x + \ell$ である．

(証明)

$$\begin{aligned}
\sum_{i=1}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i-1} &= (-\ell)\sum_{i=1}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i} \\
&= (-\ell)\left(\sum_{i=0}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i} - \binom{x}{0}(-\frac{1}{\ell})^{0}\right) \\
&= (-\ell)\sum_{i=0}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i}1^{x-i} + \ell
\end{aligned}$$

ここで，二項定理より $\sum_{i=0}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i}1^{x-i} = (1 - \frac{1}{\ell})^x$ と変形できるため，

$$\sum_{i=1}^{x}\binom{x}{i}(-\frac{1}{\ell})^{i-1} = (-\ell)(1 - \frac{1}{\ell})^x + \ell$$

という式を得る．　(Q.E.D)

定理 2.2　一様に分布する ℓ 種の値をとる x レコードのデータが持つ値の種類数 y の期待値は $f(x, \ell) = (-\ell)(1 - \frac{1}{\ell})^x + \ell$ である．

(証明)　$f(x, \ell)$ は種類数の期待値であるため，$\sum_{y=1}^{x} y \cdot Pr(y|x)$ と表せる．

$$\sum_{y=1}^{x} y \cdot Pr(y|x) = (-\ell)(1 - \frac{1}{\ell})^x + \ell \tag{3}$$

を数学的帰納法で証明する．

$x = 1$ のとき，$\sum_{y=1}^{1} y \cdot Pr(y|1) = Pr(1|1) \cdot 1 = 1$ であり，$(-\ell)(1 - \frac{1}{\ell})^1 + \ell = 1$ であるため，式 3 は成り立つ．

$x = x' - 1$ のとき，任意の ℓ で式 3 が成り立つとする．つまり，$\sum_{y=1}^{x'-1} y \cdot Pr(y|x'-1) = (-\ell)(1 - \frac{1}{\ell})^{x'-1} + \ell$ とする．このとき，$x = x'$ の場合を考えると

$$\begin{aligned}
\sum_{y=1}^{x'} Pr(y|x') \cdot y &= \sum_{y=1}^{x'}((1 - \frac{y-1}{\ell})Pr(y-1|x'-1) \\
&\quad + (\frac{y}{\ell})Pr(y|x'-1)) \cdot y \\
&= \sum_{y=1}^{x'-1}(y + 1 - \frac{y}{\ell})Pr(y|x'-1) \\
&= (1 - \frac{1}{\ell})\sum_{y=1}^{x'-1} y \cdot Pr(y|x'-1) \\
&\quad + \sum_{y=1}^{x'-1} Pr(y|x'-1) \\
&= (1 - \frac{1}{\ell})\{(-\ell)(1 - \frac{1}{\ell})^{x'-1} + \ell\} + 1 \\
&= (-\ell)(1 - \frac{1}{\ell})^{x'} + \ell
\end{aligned}$$

という式を得ることができるため，$x = x'$ のときも式 3 は成り立つ．よって，任意の $x (x \geq 1)$ について式 3 は成り立つ．　(Q.E.D)

2.3　提案モデルの分析

本節では，提案したモデルによって与えられる $Pr(y|x)$ と $f(x, \ell)$ を分析する．

レコード数 x を考えたときの y の確率分布 ($\ell = 100$) を図 2 に示す．例えば青いグラフは $Pr(y|50)$ を示しており，$y^* = 40$ のとき $Pr(40|50)$ が最大値 0.168 をとる．つまり，提案モデルに従う 50 レコードの履歴データは，全 100 種類

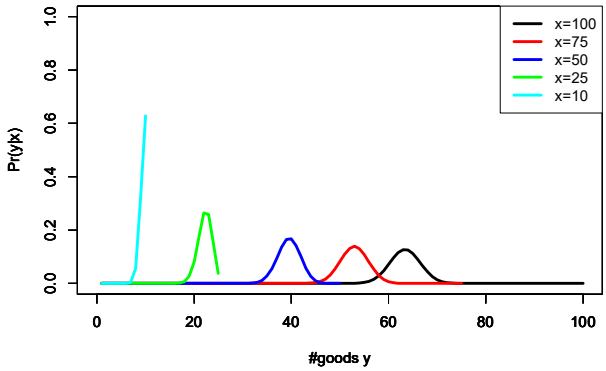

図 2　レコード数 x についての y の確率分布 ($\ell = 100$)

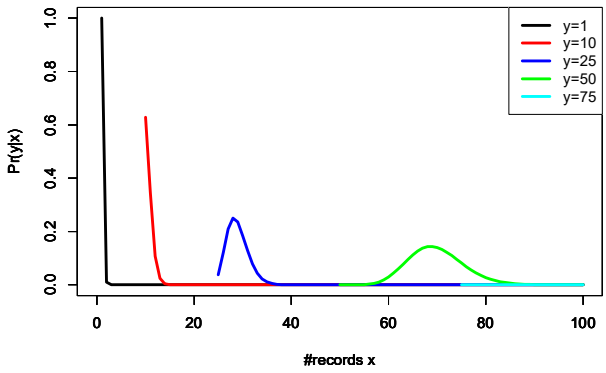

図 3　種類数 y についての x の確率分布 ($\ell = 100$)

中 40 種類の値を持つ確率が最も高いことを意味している．

また，種類数 y を与えたときの x の確率分布 ($\ell = 100$) を図 3 に示す．例えば青いグラフは $Pr(25|x)$ を示しており，$Pr(25|28)$ が最大値 (0.250) である．つまり，提案モデルに従う履歴データが 100 種類中 25 種類の項目を持つ確率は，レコード数が 28 であるとき最も高いことを意味している．

次に，ℓ を固定したときの $f(x,\ell)$ と x の関係を図 4 に示す．履歴データが x レコードあるときに，そのデータが持つ種類数 $y(\ell$ 種類中) の期待値の変化を示している．例えば黒いグラフは $f(x,100)$ のグラフであるが，これから $\ell = 100$ のとき，100 レコードの履歴データが持つ項目種類数の期待値が 63.40 であることがわかる．

3. k-匿名化に必要な疑似レコード数

2 章で提案した履歴データモデルの応用として，本稿では履歴データの k-匿名化に必要な疑似レコード数を見積もる．疑似レコード数は有用性評価指標として用いられることもあり，匿名加工データにとっては重要な値の一つである．3.1 節で記号等の定義を行い，3.2 節で疑似レコード数

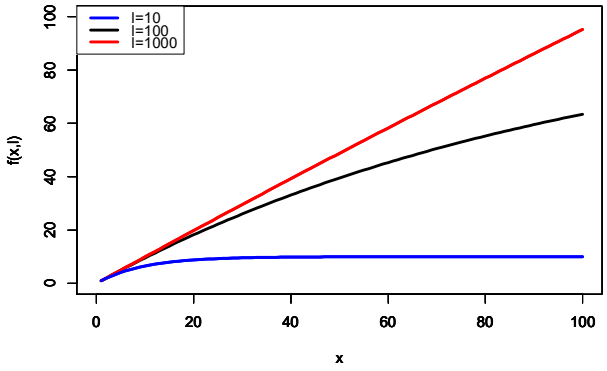

図 4　項目全種類数 ℓ の変化に伴う x と $f(x,\ell)$ の関係

表 3　T_{ex} を $c = 1$ で加工した T'_{ex}

仮名	購買商品
1	Apple
1	Book∗
2	Apple
2	Book
3	Alice∗
3	Book

の厳密解について述べ，3.3 節にて疑似レコード数の期待値を提案履歴データモデルを用いて求める．

3.1　基礎定義

定義 3.1　T を加工したデータを T' とする．T' は仮名化された T に疑似レコードを追加して，複数の仮名が同じ履歴属性の値を持つように加工されたデータである．追加された疑似レコード数を Δm とし，同じ履歴属性の値を持つ顧客のクラスタ数を c とする．i 番目のクラスタを $U_i = \{u^i_1, \ldots, u^i_{s_i}\}$ とし，s_i をクラスタ U_i の大きさ $s_i = |U_i|$(仮名数) とする．このとき，$U = U_1 \cup \cdots \cup U_c$ となる．$I(u)$ を顧客 u が持つ値の集合とし，$I(U_i) = \bigcup_{u \in U_i} I(u)$ とする．

例 3.1　T' の例として，T_{ex} を加工した T'_{ex} を表 3 に示す．T'_{ex} では Alice,Bob,Carol はそれぞれ 1,2,3 に仮名化されており，購買商品が等しいクラスタの数が 1 つになるように，2 つの疑似レコードが追加されている ($c = 1, \Delta m = 2$)．T'_{ex} 中の $*$ がついたレコードが疑似レコードである．この場合のクラスタは U_1 のみであり，$U_1 = \{1,2,3\}, s_1 = 3$ である．

3.2　疑似レコード数の厳密解

定義 3.2　クラスタ数を c とすると，疑似レコード数 Δm は

$$\Delta m = \sum_{i=1}^{c} \sum_{j=1}^{s_i} (|I(U_i)| - |I(u^i_j)|)$$

$$= \sum_{i=1}^{c} \left(s_i |I(U_i)| - \sum_{j=1}^{s_i} |I(u_j^i)| \right)$$
$$= \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$$

と表すことができる．

例 3.2 T_{ex} を $c = 1$ で加工する場合を考える．このときクラスタは $U_1 = \{1,2,3\}$ のみであり，$s_1 = 3, I(U_1) = \{Apple, Book\}, I(Alice) = \{Apple\}, I(Bob) = \{Apple, Book\}, I(Carol) = \{Book\}$ である．よって，$\Delta m = s_1|I(U_1)| - |I(u_1)| - |I(u_2)| - |I(u_3)| = 3 \cdot 2 - 1 - 2 - 1 = 2$ である．

定義 3.2 より，Δm は各クラスタの大きさ s_i，各クラスタの履歴属性の値の種類数 $|I(U_i)|$，各顧客の履歴属性の値の種類数 $|I(u_i)|$ から求めることができるが，s_i, $|I(U_i)|$, $|I(u_i)|$ の値はパラメータ c を決めて実際にデータを加工するまで手に入らないため，Δm も加工後まで求めることはできない．しかし最適な c を決めるためには，有用性評価として様々な c についての Δm を求める必要がある．

3.3 疑似レコード数の期待値

データを加工する前に Δm のおおまかな値を知ることができれば，c を決める際の指標になる．我々は定義 3.2 の s_i, $|I(U_i)|$, $|I(u_i)|$ の値を，2 章で提案した履歴データモデルを用いることによって加工前に手に入る値に置き換えることにより，Δm の期待値を計算する．本章では以下を仮定する．

定義 3.3 n 人の顧客の c 個のクラスタの大きさは全て等しく n/c である (n/c 仮定)．n 人の顧客が持つ計 m 個のレコード数は全て等しく m/n である (m/n 仮定)．

n/c 仮定と m/n 仮定より，c 個のクラスタのレコード数は等しく m/c であることがいえる．

これらの仮定より，Δm の期待値を以下のように求めることができる．

定理 3.1 顧客数 n，レコード数 m であり，全 ℓ 種類の履歴属性を持つ履歴データを，c 個のクラスタに加工するために必要な疑似レコード数の期待値 $E(\Delta m)$ は，$1/\ell$ 仮定，n/c 仮定，m/n 仮定のもと，

$$E(\Delta m) = n\ell \left((1 - \frac{1}{\ell})^{m/n} - (1 - \frac{1}{\ell})^{m/c} \right)$$

である．

(証明) 定義 3.2 より，$\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$ である．この式の s_i に n/c を，$|I(U_i)|$ に $f(m/c, \ell)$ を，$|I(u_i)|$ に $f(m/n, \ell)$ を代入することにより，以下の式を得る．

$$\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$$

$$E(\Delta m) = \sum_{i=1}^{c} \frac{n}{c} f(\frac{m}{c}, \ell) - \sum_{i=1}^{n} f(\frac{m}{n}, \ell)$$
$$= n\{(-\ell)(1 - \frac{1}{\ell})^{m/c} + \ell\}$$
$$-n\{(-\ell)(1 - \frac{1}{\ell})^{m/n} + \ell\}$$
$$= n\ell\{(1 - \frac{1}{\ell})^{m/n} - (1 - \frac{1}{\ell})^{m/c}\}$$

よって，定理 3.1 を得る． (Q.E.D)

また，定理 3.1 と k-匿名性との関係を考える．疑似レコードを追加されて c 個のクラスタに分割された履歴データが k-匿名性を満たしている場合，各クラスタの顧客数は少なくとも k である．そして定理 3.1 の履歴データにおいては，全顧客のレコード数が等しく (m/n)，全クラスタのレコード数も等しい (m/c) ため，全クラスタの顧客数が等しい (n/c) ことがわかる．つまり，$k \leq n/c$ であることが言え，定理 3.1 よりその時の疑似レコード数は以下のように求められる．

系 3.1 顧客数 n，レコード数 m であり，全 ℓ 種類の履歴属性を持つ履歴データを k-匿名化するために必要な疑似レコード数の期待値 $E'(\Delta m)$ は，

$$E'(\Delta m) \geq n\ell \left((1 - \frac{1}{\ell})^{m/n} - (1 - \frac{1}{\ell})^{km/n} \right)$$

である．

3.4 評価実験

本節では，疑似レコード数の期待値 $E(\Delta m)$ の値がどのように変化するのかを分析する．

図 5 に，$\ell = 100, c = 20$ のときのレコード数 m の変化に伴う $E(\Delta m)$ と n の関係を示す．例えば青いグラフは 10,000 レコードの履歴データ ($\ell = 100, c = 20$) についてのグラフであり，このデータの顧客数が 400 であるとき，追加する疑似レコード数の期待値が 30,850 であり，顧客数が増えるほど必要な疑似レコード数も増えることがわかる．

また，$n = 400, c = 20$ のときの全種類数 ℓ の変化に伴う $E(\Delta m)$ と m の関係を図 6 に示す．例えば赤いグラフは $\ell = 1,000, n = 400, c = 20$ のときのグラフであるが，$m = 63,037$ のときに極大 (324,570 レコード) であり，そこからは $E(\Delta m)$ は減少している．このことから，履歴データのレコード数が莫大になると，必要な疑似レコード数は減少することがわかる．

次に，クラスタ数 c と $E(\Delta m)$ の関係を図 7 に示す．クラスタ数が増えるほど必要な疑似レコード数は減少する．また，k-匿名化に必要な疑似レコード数の期待値 $E'(\Delta m)$ の最小値と k の関係を図 8 に示す．このとき，$n = 400, m = 38,000, \ell = 2,700$ である．k の値を大きくするほど $E'(\Delta m)$ の最小値も大きくなり，加工データの有用性が下がる．

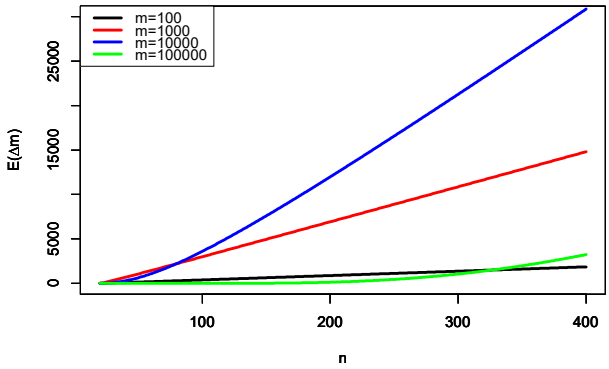

図 5　レコード数 m の変化に伴う $E(\Delta m)$ と n の関係 $(\ell = 100, c = 20)$

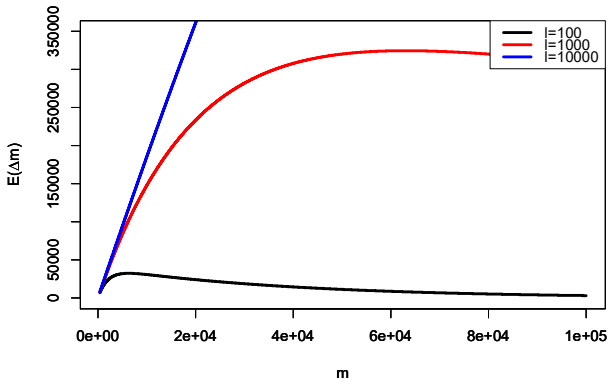

図 6　全種類数 l の変化に伴う $E(\Delta m)$ と m の関係 $(n = 400, c = 20)$

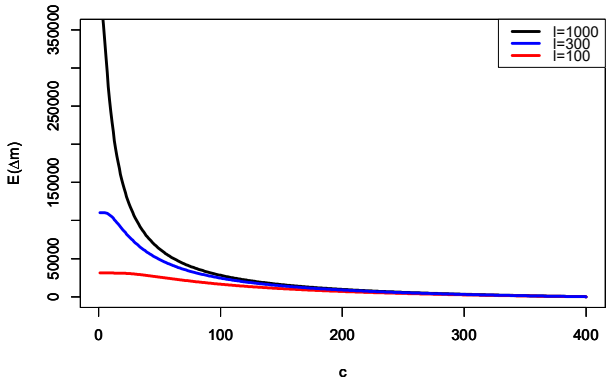

図 7　全種類数 l の変化に伴う $E(\Delta m)$ と c の関係 $(n = 400, m = 10,000)$

最後に，提案モデルでは履歴データ中で ℓ 種類の履歴属性の値が全て等確率 $(1/\ell)$ で生起することを仮定しているが，この仮定の妥当性についても分析する．PWSCUP 2016[5] で用いられた，400 人分の購買履歴データ (Online retail Dataset[7]) についての購買商品の頻度分布を図 9 に示す．このデータには 2,781 種類の購買商品が登場するが，最も多く購買されている商品の発生頻度が 1,000 回以上であるのに対し，最も購買されていない商品の発生頻度は 1 回である．項目によって発生確率に大きな差がある履歴データは多く，$1/\ell$ 仮定は妥当なものではないことがわかる．

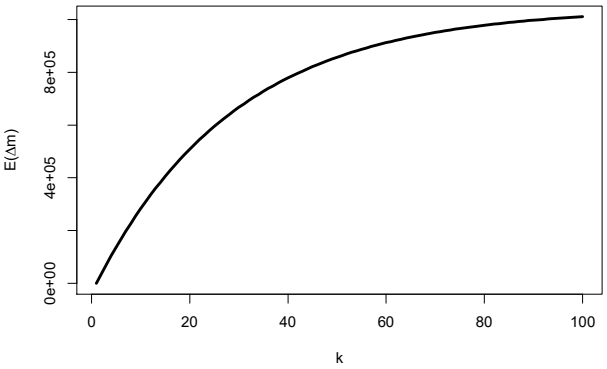

図 8　$E'(\Delta m)$ の最小値と k の関係

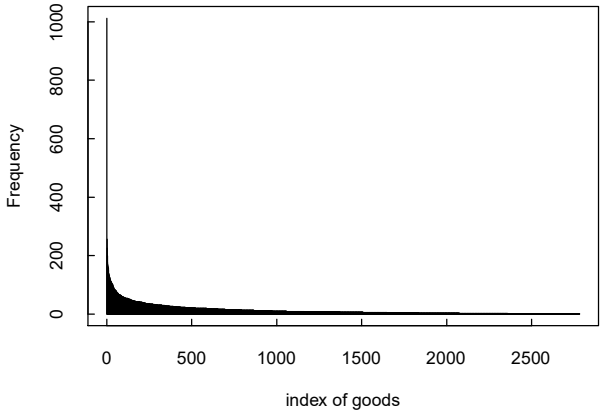

図 9　400 人分の購買履歴データの購買商品の頻度分布

4. おわりに

本稿では，履歴データ中のある属性の値が全て等確率で生起する仮定のもと，履歴データ中に登場する項目の種類数の確率分布とその期待値を与える数理モデルを提案した．提案モデルによって，「x レコードの履歴データが持つ値 (全 ℓ 種類) の種類数 y の期待値」を求めることができる．

また，提案モデルを応用することにより，履歴データを k-匿名化するために必要な疑似レコード数の期待値を，元データの統計量やパラメータ k から求めた．提案モデルを応用することにより，疑似レコード数以外のデータ評価値も加工前に求めることができることが期待できる．しかしながら，本研究の仮定 $(1/\ell$ 仮定，n/c 仮定，m/n 仮定) は強く妥当ではないため，データに応じた改善が必要である．

履歴データモデルの改善と，他の加工データ評価値への

提案モデルの適用を今後の課題とする．

参考文献

[1] 村上 隆夫，荒井 ひろみ，井口 誠，小栗 秀暢，菊池 浩明，黒政 敦史，中川 裕志，中村 優一，西山 賢志郎，野島 良，波多野 卓磨，濱田 浩気，山岡 裕司，山口 高康，山田 明，渡辺 知恵美，“PWS Cup 2019: ID 識別・トレース推定に強い位置情報の匿名加工技術を競う”，コンピューターセキュリティシンポジウム (CSS 2019), pp. 1485–1492. (2019)

[2] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp.557–570. (2006)

[3] 濱田 浩気，荒井 ひろみ，小栗 秀暢，菊池 浩明，黒政 敦史，中川 裕志，西山 賢志郎，波多野 卓磨，村上 隆夫，山岡 裕司，山田 明，渡辺 知恵美，“PWS Cup 2018: 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～”, コンピューターセキュリティシンポジウム (CSS 2018), pp. 935–940. (2018)

[4] 原田 玲央、伊藤 聡志、菊池 浩明，“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”, 暗号と情報セキュリティシンポジウム (SCIS 2017), pp. 1–8. (2017)

[5] 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匿名加工せよ”, コンピューターセキュリティシンポジウム (CSS 2016), pp. 271–278. (2018)

[6] Michael Mitzenmacher, Eli Upfal, “Probability and Conputing: Randomized Algolithms and Probabilistic Analysis”, Cambridge University Press, pp. 32–34. (2005)

[7] Online Retail Data Set, `https://archive.ics.uci.edu/ml/datasets/online+retail`, December 17, 2018.