Vol.2019-DPS-178 No.31
Vol.2019-CSEC-84 No.31
2019/3/5

背景知識の違いによる匿名加工データの攻撃者モデルの分類と評価

伊藤 聡志[1]　菊池 浩明[1]

概要：匿名加工は，購買履歴データのような元データから個人が識別されることを防ぐために，個人識別情報を加工する技術である．データを匿名加工する際には，データを悪用しようとする攻撃者を想定し，リスクを評価する必要がある．しかしながら，データに対する攻撃者をどう想定したらよいかはいまだ不明である．本研究では，履歴データのある属性から背景知識を得る攻撃者を想定し，攻撃者の持つ背景知識に当てはまるレコード数とユーザ数を用い，データリスク評価の理論的なモデルを提案する．また，提案したモデルを用いて実際の履歴データのリスク評価実験を行う．

キーワード：個人情報，プライバシーリスク評価，匿名加工，履歴データ

Satoshi Ito[1]　Hiroaki Kikuchi[1]

1. はじめに

匿名加工は，元データから個人が識別されることを防ぐために，個人識別情報を加工する技術である．企業や組織は収集したビッグデータを活用する際，そのデータ内の個人が再識別されるリスクを評価し，匿名加工することを求められる．一方，そのようなデータから個人を識別しようとする攻撃者は，公開されている匿名加工データだけでなく，追加の背景知識を用いることが予想される．しかしながら，データについてのどのような背景知識が危険であるのか，どういった攻撃者が危険であるのかは不明であった．加えて，購買履歴データのようなデータを匿名加工する際に，どの属性（列）を加工したらよいのかを判断するための指標も不明であった．

Domingo-Ferrer らはデータセットに対する攻撃者想定として，最大知識攻撃者モデルを提案した [1]．最大知識攻撃者モデルでは，攻撃者は元のデータセットと匿名加工されたデータの両方のすべてを背景知識として持っていることを想定されている．しかしながら，この攻撃者想定はあまりにも協力であり，現実的ではない．また El Eman らは，データセットに対して現実世界で実行されうる 4 種類の攻撃（故意の攻撃，故意でない攻撃，データ侵害，公開データ）を想定し，それらのリスクを測定した [2]．しかしながら，これらのリスク評価には「攻撃が行われる確率」といった主観的な値や，「データ侵害が発生する確率」といった時間や場合によって変化する値が用いられており，求めるのが困難であった．

本研究の目的は，現実的な攻撃者の想定とデータセットのリスク評価である．我々はレコードと属性によって構成される履歴データに注目し，ある属性から背景知識を得る攻撃者を想定する．攻撃者の持つ背景知識に当てはまるレコード数と顧客数を用い，データの危険度の理論的なモデルを提案する．また，提案したモデルを評価するために，公開データセットを用いた実験を行う．本研究の貢献は以下の通りである．

(1) 履歴データの一部を背景知識として持つ攻撃者の危険度を近似する理論的なモデルの提案
(2) 匿名加工の際に加工するべき属性を判断するための指標の提案
(3) 実際のデータに対するリスク評価実験

2. 基礎定義

2.1 データモデル

本研究では，レコード（行）と属性（列）によって構成され，個人を表す識別子を持つ履歴データを研究する．記号等を以下のように定義する．

定義 2.1　履歴データを T とし，T のレコード数を m，

[1] Meiji University, Nakano, Tokyo, 4-21-1

Vol.2019-DPS-178 No.31
Vol.2019-CSEC-84 No.31
2019/3/5

表 1　履歴データ T の例 T_{example}

User ID	Date	Time	Goods	Price	Number
1	2010/12/1	8:45	Bread	1.45	2
1	2010/12/1	8:45	Book	3.75	1
1	2010/12/1	20:10	Tea	0.85	2
2	2010/12/1	10:03	Bread	1.45	3
1	2010/12/2	15:07	Tea	0.85	3
3	2010/12/2	11:57	Bread	1.45	4
3	2010/12/2	11:57	Juice	1.25	4
3	2010/12/3	15:54	Book	3.75	1
3	2010/12/3	15:54	Tea	0.85	10
3	2010/12/3	15:54	Juice	1.45	10

ユーザ数を n とする．履歴 T の属性 X の取りうる値の集合を D_X とし，T における X のユニークな値の数を ω_X とする．すなわち，$\omega_X = |D_X|$ である．D_X の要素 x について，履歴 T で x を満たすレコード（行）インデックスの集合を R_x とし，x を満たすユーザの集合を U_x とする．T を匿名化してユーザ ID を仮名化した匿名化データを T' とする．

例 2.1　T の例として，3 人のユーザ（ユーザ 1,2,3）の 3 日間（2010/12/1〜2010/12/3）の購買履歴データ T_{example} を表 1 に示す．例えば，仮名 2 は 2010/12/1 にパンを購入していることがわかる．T_{example} は $m = 10, n = 3$ の履歴データであり，X=Date のとき，$D_X = \{2010/12/1, 2010/12/2, 2010/12/3\}$, $\omega_X = 3$ である．また，$x = 2010/12/1$ のとき，$R_x = \{1,2,3,4\}, U_x = \{1,2\}$ である．

2.2　攻撃者モデル

本研究では，攻撃者が履歴 T に属するユーザ u の属性 X についての背景知識 x を偶然得ることを想定する．

定義 2.2　攻撃者が背景知識 x を得る確率 $Pr(x)$ は，x の T における頻度に比例する，すなわち，$Pr(x) = |R_x|/m$ である．また，T のレコード数 m と属性 X の種類数 ω_X は与えられているものとする．

匿名化データ T' を与えられた攻撃者は，背景知識として x を含む T のレコードにアクセスできるとき，対応する T' の仮名の真のユーザの候補として U_x を得る．従って，再識別を表す事象 idf が生起するリスクを，x の条件付確率として次のように定める．

定義 2.3　攻撃者が背景知識 x から個人を識別 (idf) する条件付き確率 $Pr(\text{idf}|x)$ を $Pr(\text{idf}|x) = 1/|U_x|$ とする．

定義 2.2,2.3 より，攻撃者が背景知識 x を得ることと，攻撃者が背景知識 x から個人を識別することの同時確率 $Pr(\text{idf}, x)$ は，

$$Pr(\text{idf}, x) = Pr(x)Pr(\text{idf}|x) = \frac{|R_x|}{m}\frac{1}{|U_x|}$$

である．また，ここで $|R_x|/|U_x| = \alpha_x$ とおくと，

$$Pr(\text{idf}, x) = \frac{\alpha_x}{m}$$

とも表せる．α_x は x についてのユーザ当たりの平均レコード数 [レコード/人] を意味しており，本論文の解析に重要な役割を果たす．そこで，これを次のように定義する．

定義 2.4　背景知識 x による平均レコード数を α_x とする．属性 X における α_x の平均を α_X と表し，$\alpha_X = \frac{1}{\omega_X}\sum_{x\in D_X}\alpha_x$ とする．

例 2.2　T_{example} の Date 属性についての x, $|R_x|$, $Pr(x)$, $|U_x|$, $Pr(\text{idf}|x)$, $Pr(\text{idf}, x)$ を表 2 に示す．T_{example} の Date 属性の場合，$D_X = \{2010/12/1, 2010/12/2, 2010/12/3\}$ である．攻撃者が背景知識 x =2010/12/3 を得る確率は，$R_x = \{8,9,10\}$ であるため $Pr(x) = 3/10$ であり，その背景知識からユーザ u を識別できる確率は，$U_x = \{3\}$ なので，$Pr(\text{idf}|x) = 1/1$ となる．この場合，攻撃者が背景知識 x によって u を識別できる確率は $Pr(\text{idf}, x) = Pr(x)Pr(\text{idf}|x) = 0.3 \cdot 1 = 0.3$ である．または，$\alpha_x = 3/1 = 3$ であるので，

$$Pr(\text{idf}, x) = \frac{\alpha_x}{m} = \frac{3}{10} = 0.3$$

である．

2.3　リスクモデル

本研究では以下に定義する平均識別確率 $Pr(\text{idf}, X)$ を，履歴 T の属性 X に関する危険度とする．

定義 2.5　(平均識別確率) 履歴 T の属性 X のある値を背景知識 x として与えられた攻撃者により，あるユーザ u が識別される確率 $Pr(\text{idf}|x)$ の期待値を，属性 X の平均識別確率 $Pr(\text{idf}, X)$ とする．

定義 2.4 より，

$$Pr(\text{idf}, X) = \sum_{x\in D_X} Pr(\text{idf}, x) = \sum_{x\in D_X}\frac{\alpha_x}{m}$$

である．

例 2.3　X =Date の場合，T_{example} の属性 X から背景知識 x を得た攻撃者の平均識別確率は

$$Pr(\text{idf}, X) = \sum_{x\in D_X}\frac{\alpha_x}{m} = \frac{2+1.5+3}{10} = 0.65$$

である．これは，攻撃者が T_{example} の Date 属性からあるユーザ u の背景知識を得たとき，u を平均 65% の確率で識別できることを意味する．

また，リスクの計算コストを以下のように定義する．

定義 2.6　リスク計算のコストは，計算に用いるレコード数に比例する．

例 2.4　履歴 T_{example} の全レコードの Price 属性の平均値を求める場合，計算コストは 10 である．また，2010/12/1 の Price 属性の平均値を求める場合，計算コストは 4 である．

Vol.2019-DPS-178 No.31
Vol.2019-CSEC-84 No.31
2019/3/5

表 2　T_{example} の Date 属性に対する攻撃者の識別確率

x	$\lvert R_x\rvert$	$Pr(x)$	$\lvert U_x\rvert$	$Pr(\text{idf}\lvert x)$	$Pr(\text{idf}, x)$
2010/12/1	4	0.4	2	0.5	0.2
2010/12/2	3	0.3	2	0.5	0.15
2010/12/3	3	0.3	1	1	0.3
合計	10	1.0			0.65

3. リスク近似モデルの提案

平均識別確率を求めるためには，定義 2.5 より，履歴 T の属性 X に出現するすべての x について，α_x を求める必要がある．しかしながら，ビッグデータに対してすべての α_x を計算するのは困難であるため，これを近似する方法を検討する．平均識別確率を計算するモデルとして，本章では以下の 3 つを提案する．

(1) 平均モデル
(2) 最小コストモデル
(3) サンプリングモデル

3.1 厳密解

匿名化データ T' の再識別のリスクは，攻撃者に与えられる背景知識の属性 X に依存して決まる．そこで，X を与えられた時の再識別リスク $R(X)$ を，属性 X の平均識別確率と定める．すなわち，$R(X) = Pr(\text{idf}, X)$ とする．$R(X)$ の厳密解を求めるためには，履歴 T の属性 X に出現するすべての x について α_x を求める必要があるため，この場合の計算コストは m である．

3.2 平均モデル

平均モデルは，属性 X のリスクを α_x の平均 α_X を用いて求めるモデルである．以下のように定義を行う．

定義 3.1　平均モデルによって求められる属性 X のリスクを $R_{mean}(X)$ で示し，

$$R_{mean}(X) = \frac{\alpha_X \omega_X}{m}$$

とする．

平均レコード数の平均 α_X で定めた平均モデルのリスクは次のように厳密解を与えている．

定理 3.1　$R_{mean}(X)$ を定義 3.1 による，平均モデルによって求められるリスクとする．このとき，$R_{mean}(X) = Pr(\text{idf}, X)$ である．

(Proof)　定義 3.1,2.4 より，

$$\begin{aligned} R_{mean}(X) &= \frac{\alpha_X \omega_X}{m} \\ &= \frac{\omega_X}{m} \sum_{x \in D_X} \frac{\alpha_x}{\omega_X} \\ &= \sum_{x \in D_X} \frac{\alpha_x}{m} \\ &= \sum_{x \in D_X} Pr(x) Pr(\text{idf}|x) \\ &= Pr(\text{idf}, X) \end{aligned}$$

であり，定理 3.1 を得る．　(Q.E.D)

例 3.1　T_{example} の Date 属性の場合，$\alpha_X = (2 + 1.5 + 3)/3 = 13/6$ であるため，

$$R_{mean}(X) = \frac{\alpha_X \omega_X}{m} = \frac{13/6 \cdot 3}{10} = 0.65$$

である．

このモデルでは α_X を求める際に，履歴 T の属性 X に出現するすべての x について α_x を計算する必要があるため，この場合の計算コストは m である．

3.3 最小コストモデル

最小コストモデルは，全ての x について $\alpha_x = 1$ と近似して，属性 X のリスクを最小の計算コストで求めるモデルである．以下のように定義を行う．

定義 3.2　最小コストモデルによる属性 X のリスクを，

$$R_{cost}(X) = \frac{\omega_X}{m}$$

とする．

例 3.2　T_{example} の Date 属性の場合，

$$R_{cost}(X) = \frac{\omega_X}{m} = \frac{3}{10} = 0.3$$

である．

定理 3.2　最小コストモデルの誤差率は $|\frac{1}{\alpha_X} - 1|$ である．

(Proof)　定理 3.1 により $Pr(\text{idf}, X) = \alpha_X \omega_X / m$ を用いると，$R_{cost}(X)$ の厳密解に対する誤差率は，

$$\begin{aligned} &= \frac{|R_{cost}(X) - Pr(\text{idf}, X)|}{Pr(\text{idf}, X)} \\ &= \frac{|\frac{\omega_X}{m} - \frac{\alpha_X \omega_X}{m}|}{\frac{\alpha_X \omega_X}{m}} = |\frac{1}{\alpha_X} - 1| \end{aligned}$$

となるため，定理 3.2 を得る．　(Q.E.D)

定義 2.2 より，T のレコード数 m と属性 X の種類数 ω_X は与えられている情報であり，このモデルでは履歴 T のレコードを用いて α_X 等を計算する必要が無いため，計算コストは 0 である．

3.4 サンプリングモデル

サンプリングモデルは，D_X からランダムに選んだ複数個の要素についての α_x を求め，これの平均を属性 X の平均レコード数 α_X の近似値であるとして属性 X のリスクを求めるモデルである．このとき，サンプリングするのは 1 つのレコードではなく，D_X からランダムに選んだ複数個の要素を満たすすべてのレコードであることを注意せよ．例えば，T_{example} の Date 属性のうち “2010/12/1” がランダムに選ばれた場合，T_{example} からこれを満たすレコード（この場合 4 レコード）をすべてサンプリングする．以下

Vol.2019-DPS-178 No.31
Vol.2019-CSEC-84 No.31
2019/3/5

表 3　提案モデルの概要

Model	Risk	Error Rate	Cost
Exact Solution	$R(X)$	0	m
Mean	$R_{mean}(X)$	0	m
Low Cost	$R_{cost}(X)$	$\frac{1}{\alpha_X}-1$	0
Sampling	$R_{sample}(X)$	Eq. (1)	sm/ω_X

のように定義を行う．

定義 3.3　s をサンプリング数とし，$D'_X=\{x_1,\ldots,x_s\}$ を D_X からランダムにサンプリングされた，要素が s 個の部分集合とする．このとき，$\alpha_{x'}=\frac{1}{s}\sum_{i=1}^{s}\alpha_{x_i}$ とする．最小コストモデルによる属性 X のリスク $R_{sample}(X)$ を，

$$R_{sample}(X)=\frac{\alpha_{x'}\omega_X}{m}$$

とする．また，σ_s を s 個のサンプルの，$R_{sample}(X)$ についての標準偏差とする．

例 3.3　T_{example} の X =Date 属性の場合，$s=2$, $D'_X=\{2010/12/1, 2010/12/3\}$ とすると，$\alpha_{x_1}=2$, $\alpha_{x_2}=3$ であるため，

$$R_{sample}(X)=\frac{\alpha_{x'}\omega_X}{m}=\frac{2.5\cdot 3}{10}=0.75$$

である．

定理 3.3　サンプリングモデルの誤差率の最大値は

$$\frac{\sigma_s m}{\sqrt{|s|}\omega_X\alpha_X} \tag{1}$$

である．

(Proof)　定理 3.1 より，$Pr(\text{idf},X)=\alpha_X\omega_X/m$ である．このとき，90%信頼区間を仮定すると，$Pr(\text{idf},X)$ との絶対誤差は $|R_{sample}(X)-Pr(\text{idf},X)|<Var[Pr(\text{idf},X)]=\sigma_s/\sqrt{s}$ となる．よって，$R_{sample}(X)$ と厳密解の相対誤差率は

$$=\frac{|R_{sample}(X)-Pr(\text{idf},X)|}{Pr(\text{idf},X)}$$

$$=\frac{\frac{1}{\sqrt{s}}\sigma_s}{\frac{\alpha_X\omega_X}{m}}=\frac{\sigma_s m}{\sqrt{|s|}\omega_X\alpha_X}$$

となるため，定理 3.3 を得る．　(Q.E.D)

このモデルでの $\alpha_{x'}$ の計算コストは，D'_X の要素が $1/\omega_X$ で一様に選ばれるならば，これは $p=\frac{1}{\omega_X}$, 期待値 $\mu=\frac{m}{\omega_X}$ の二項分布であるため，sm/ω_X である．

表 3 に各モデルの概要をまとめる．厳密解と平均モデルの計算コストは最大（m）であるが，誤差はゼロである．一方，最小コストモデルのコストはゼロであるが，誤差は大きくなる．サンプリングモデルの計算コストと誤差はサンプリングサイズ s に依存し，s が大きくなるほど誤差は小さくなり，計算コストは大きくなる．

4. 評価実験

4.1 実験目的

前節で提案したモデルを用いて，実際のデータに対するリスク評価実験を行う．実験のために，UCI Machine Learning Repository[3] より公開されている以下の 3 つのデータセットを用いる．

(1) T_1: Online Retail Data Set [4]
(2) T_2: Diabetes Data Set [5]
(3) T_3: Adult Data Set [6]

T_1,T_2,T_3 はそれぞれ，英国の 1 年間の購買履歴データ，10 年間の糖尿病患者・入院データ，国税調査による所得データである．各データの m,n, 属性数を表 4 に示す．

表 4　公開データセット T_1,T_2,T_3 の統計量

	m	n	属性数
T_1	38,087	400	7
T_2	101,766	71,518	50
T_3	32,561	32,561	16

4.2 データセットの分析

表 5 に T_1 の各属性の概要を示す．このデータは 7 属性から成るデータであるが，本研究ではユーザ ID・伝票 ID を除いた 5 属性を X の候補として用いる．各属性の α_x の分布を図 1〜5 に示し，各属性についての α_X と ω_X を表 6 に示す．

Date,Time 属性はユーザごとの平均レコード数 α_x が大きく，100 レコードを超える x もあり，例えば 2011/8/28 には 1 人のユーザが 122 レコードの購買をしている．本研究ではこういった x は，背景知識として得やすく，これを得た攻撃者が個人を識別しやすいので，大変危険であると評価される．一方，Goods,Price,Number 属性では α_x が小さく，多くの x について $\alpha_x=1$ である．x 軸を $|U_x|$，y 軸を $|R_x|$ とした，T_1 の Date,Price 属性についての散布図を図 6,7 に示す．赤直線は $y=\alpha_X\cdot x$（平均モデル）を示し，緑直線は $y=x$（最小コストモデル）を示す．

表 7 に T_2 の各属性の概要を示す．このデータは 50 属性から成るデータであるが，本研究ではそのうち，攻撃者が背景知識として得ることが想定される 4 属性に注目する．表 8 に T_3 の各属性の概要を示す．このデータは 17 属性から成るデータであるが，本研究ではそのうち，攻撃者が背景知識として得ることが想定される 4 属性に注目する．表 6 に T_2,T_3 の各属性についての α_X と ω_X を示し，T_2 の Age,Time 属性の α_x の分布を図 8 に示す．T_3 は $m=n$ より，任意の x で $|R_x|=|U_x|$ であるため，$\alpha_X=1$ である．

4.3 提案モデルの精度と計算コスト

T_1,T_2,T_3 の各属性の危険度を平均モデル，最小コストモデル，サンプリングモデルによって効率よく求める．表 9 に各モデルの評価値を示す．定理 3.1 により，平均モデルによる評価値 $R_{mean}(X)$ は表 6 の $R(X)$ と一致する．サンプリングモデルによる評価値 $R_{sample}(X)$ は，$s=10$ の

Vol.2019-DPS-178 No.31
Vol.2019-CSEC-84 No.31
2019/3/5

表 5　T_1 の概要

Attribute	Detail
User ID	ID of user (5 digit number)
Receipt ID	ID of receipt (6 digit number)
Date	Purchase date (yyyy/mm/dd)
Time	Purchase time (hh:mm)
Goods	ID of purchased goods (number and character)
Price	Price of purchased goods (Pound sterling)
Number	Quantity of purchased goods (number)

表 6　T_1, T_2, T_3 の分析結果

T	X	α_X	ω_X	$Pr(\mathrm{idf}, X)$	σ
T_1	Time	22.23	551	0.322	0.228
	Date	24.42	290	0.186	0.140
	Goods	1.32	2781	0.097	0.151
	Price	2.49	184	0.012	0.066
	Number	3.15	97	0.008	0.043
T_2	Days	1.05	14	$1.45 \cdot 10^{-4}$	$1.66 \cdot 10^{-4}$
	Age	1.35	10	$1.33 \cdot 10^{-4}$	$3.20 \cdot 10^{-4}$
	Race	1.31	6	$7.73 \cdot 10^{-5}$	$2.08 \cdot 10^{-4}$
	Gender	1.28	3	$3.78 \cdot 10^{-5}$	$1.81 \cdot 10^{-3}$
T_3	Age	1	73	$2.24 \cdot 10^{-3}$	$1.01 \cdot 10^{-2}$
	Occupation	1	15	$4.61 \cdot 10^{-4}$	$1.21 \cdot 10^{-3}$
	Martial	1	7	$2.15 \cdot 10^{-4}$	$1.20 \cdot 10^{-3}$
	Race	1	5	$1.54 \cdot 10^{-4}$	$4.79 \cdot 10^{-4}$

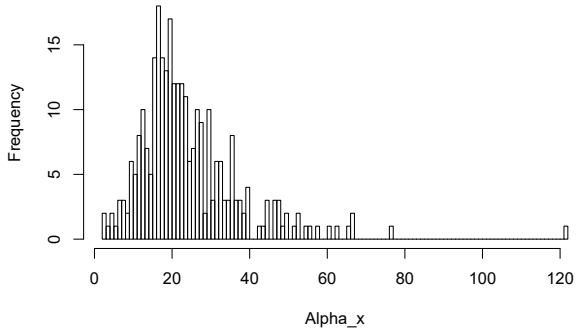

図 1　T_1 の X =Date 属性についての α_x の分布

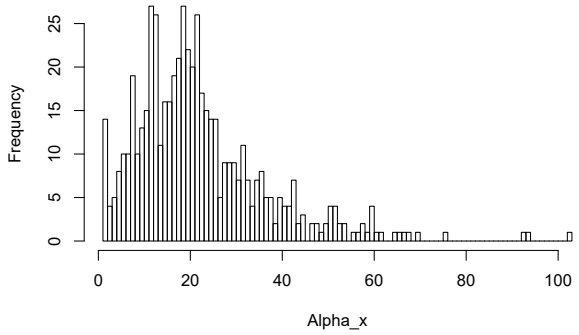

図 2　T_1 の X =Time 属性についての α_x の分布

ときの 90%($\mu \pm \sigma$) の信頼区間を示している．表中の*印がついている値は，そのデータで最も危険であると評価された属性のリスクである．例えば T_1 について，平均モデル（=厳密解）では Time 属性が最も危険であると評価されて

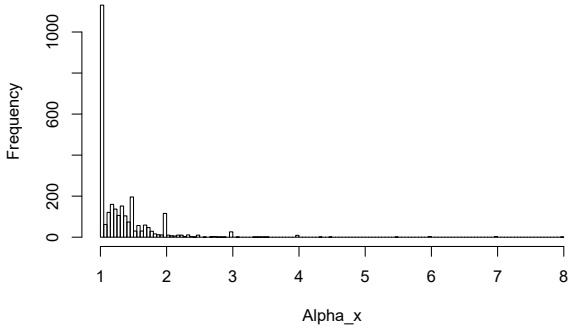

図 3　T_1 の X =Goods 属性についての α_x の分布

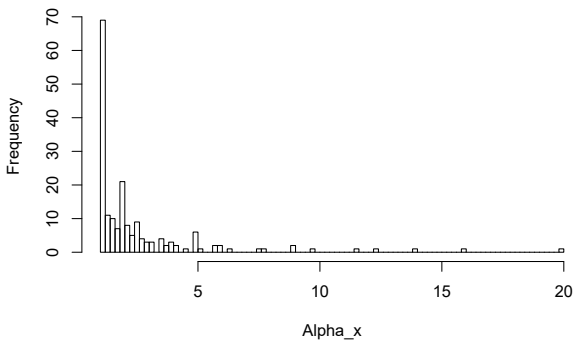

図 4　T_1 の X =Price 属性についての α_x の分布

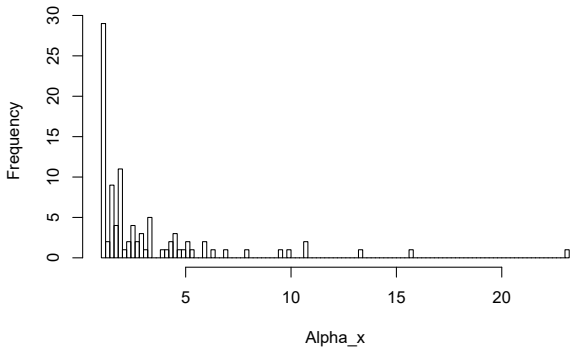

図 5　T_1 の X =Number 属性についての α_x の分布

表 7　T_2 の概要

Attribute	Detail
Patient ID	ID of patient
Race	Race of patient
Gender	Gender of patient
Age	Age of patient
Time	Time in hospital

表 8　T_3 の概要

Attribute	Detail
ID	ID of user
Age	Age of user
Marital	Marital-status of user
Occupation	Occupation of user
Race	Race of user

いるのに対し，最小コストモデルでは Goods 属性が最も危険であると評価されている．サンプリングモデルにおい

Vol.2019-DPS-178 No.31
Vol.2019-CSEC-84 No.31
2019/3/5

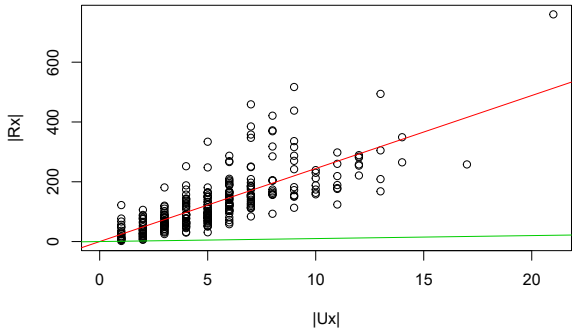

図 6　T_1 の X =Date 属性についての散布図

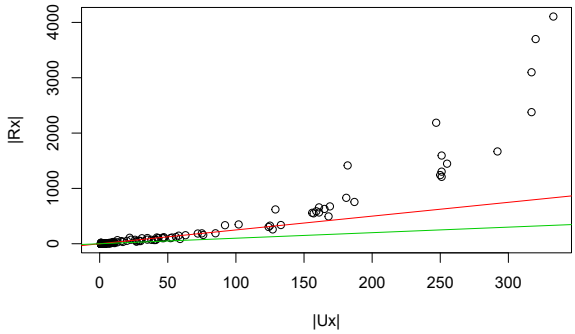

図 7　T_1 の X =Price 属性についての散布図

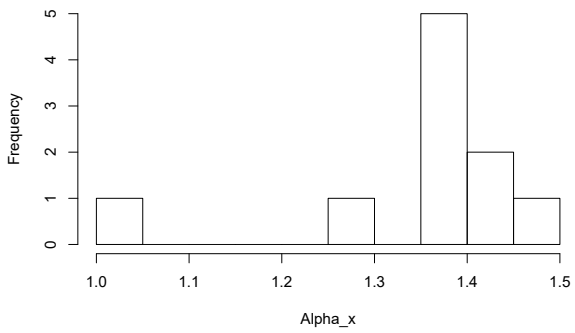

図 8　T_2 の X =Age 属性についての α_x の分布

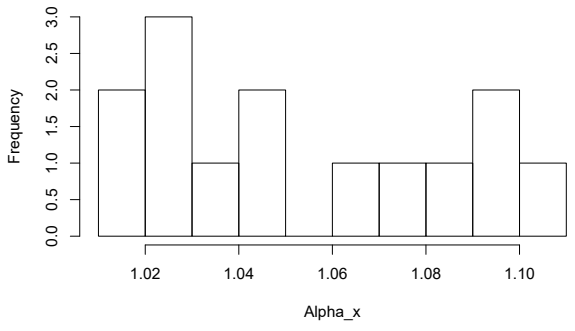

図 9　T_2 の X =Time 属性についての α_x の分布

ては，信頼区間の半順序関係における極大値となる属性は Time であった．

表 10 に各モデルのコストと誤差の値を示し，図 10 に T_1 の Date 属性についての，各モデルの計算コストと誤差の散布図を示す．X 軸は計算コスト（レコード数）の対数であり，Y 軸は厳密解 $Pr(\mathrm{idf}, X)$ との絶対誤差である．図中

表 9　各モデルによって近似された平均識別確率

T	X	$R_{mean}(X)$	$R_{cost}(X)$	$R_{sample}(X)(s=10)$
T_1	Time	*0.3217	0.0145	*[0.1411, 0.5998]
	Date	0.1860	0.0076	[0.1267, 0.2786]
	Goods	0.0965	*0.0730	[0.0718, 0.0982]
	Price	0.0121	0.0048	[0.0036, 0.0132]
	Num	0.0080	0.0025	[0.0017, 0.0152]
T_2	Days	*1.45E-04	*1.38E-04	*[1.46E-04, 1.52E-04]
	Age	1.33E-04	9.83E-05	[1.21E-04 , 1.42E-04]
	Race	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	Gender	3.78E-05	2.95E-05	[3.08E-05 , 4.30E-05]
T_3	Age	*2.24E-03	*2.24E-03	*[2.24E-03, 2.24E-03]
	Occupation	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	Martial	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	Race	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]

表 10　各モデルのコストと誤差

Model	Cost	Error
Mean	38087	0
Sample	131.3	0.073
Cost	0	0.178

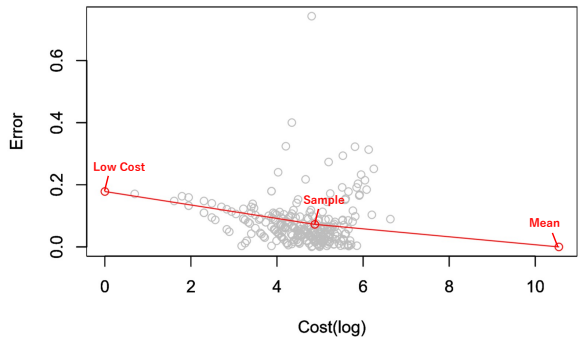

図 10　各モデルのコストと誤差の散布図

の赤い点がこれらのモデルの結果を表している．灰色の点は D_X の ω_X 個の要素のリスク評価結果を示しており，それらの重心をサンプリングモデルの代表の点としている．サンプリングモデルはこれらの ω_X 個の点から s 個をランダムに選んでリスク評価をすることに注意せよ．

T_1 の Date 属性の D_X から 50 種類の x を 1000 回ランダムサンプリングしたときの α_X の分布を図 11 に示す．また，サンプリング種類数毎の α_X の平均と標準偏差を表 11 に示す．これらの図表からわかるように，属性 X からランダムな x を選び，それについての α_x を求めることで，α_X を近似することができる．サンプリングのサイズに応じて，急速に真値に収束していることがわかる．これにより，本章ではサンプリングサイズ 10 でリスク評価を行った．

5. まとめ

本稿では，履歴データのある属性から背景知識を得る攻撃者を想定し，その平均識別確率を用いてデータのリスク評価を行うモデルを提案した．また，平均識別確率を近似する 3 つのモデルを提案し，それらを用いて購買履歴データ，入院記録データ，世帯収入データの 3 つの実際のデータのリスク評価を行い，どの属性が危険であるのかを評価

Vol.2019-DPS-178 No.31
Vol.2019-CSEC-84 No.31
2019/3/5

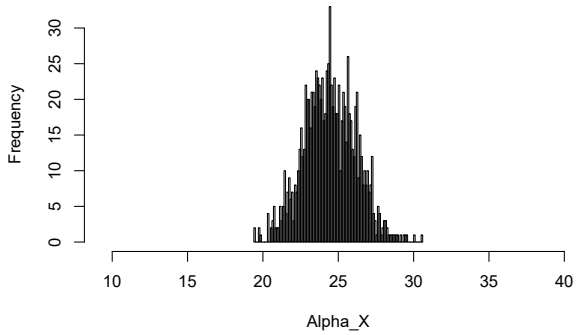

図 11 T_1 の Date 属性から 50 グループをサンプリングしたときの α_X の分布

表 11 T_1 の Date 属性から複数グループをサンプリングしたときの α_X の平均

#Sample	平均	標準偏差
1	24.03	13.33
50	24.47	1.75
100	24.39	1.09
150	24.41	0.78
200	24.41	0.53
250	24.42	0.33
ω_X	24.42	0

した．匿名加工をする際にこのリスク評価モデルを用いることによって，どの属性を加工・削除するか？等の加工指針を立てることができる．

参考文献

[1] Josep Domingo-Ferrer, Sara Ricci and Jordi Soria-Comas, “Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker”, 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), *IEEE*, 2015.

[2] Khaled El Emam, Luk Arbuckle, “Anonymizing Health Data Case Studies and Methods to Get You Started”, *O’Reilly*, 2013.

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

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

[5] Diabetes 130-US hospitals for years 1999-2008 Data Set , `https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008`, December 17, 2018.

[6] Adult Data Set , `https://archive.ics.uci.edu/ml/datasets/adult`, December 17, 2018.