明治大学大学院　先端数理科学研究科

２０１８年度

修士学位請求論文

背景知識の違いによる匿名加工データへの攻撃者モデルの評価

学位請求者　先端メディアサイエンス専攻

伊藤　聡志

あらまし

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．

目 次

第1章　はじめに

1.1　研究背景

個人情報データや購買履歴データ等のビッグデータは非常に有用であり，利活用することによって大きな利益を生むことができる．例えば，Google, Amazon, Facebook, Apple (GAFA) に代表されるターゲット広告では，GAFA が収集した個人データをビジネスに最大限に利用して巨額の利益を得ている．しかし，それらのビッグデータの活用にはプライバシーの課題がある．例えば Sweeney らは匿名加工されたデータから郵便番号・性別・誕生日の 3 属性の情報によって，アメリカ合衆国の人口の 87 %が一意に特定されることを示している [1]．また日本では，2013 年には JR 東日本が Suica のデータを他者に販売しようとして顧客から大きな批判を受けた事件 [2] を受けて，個人情報保護法 [3] が 2016 年に改正され,「匿名加工情報」という概念が定義された．

1.2　匿名加工・再識別について

企業や組織は収集したビッグデータを利活用する際，そのデータのプライバシーリスクを評価し，適切なレベルの匿名加工を行う必要がある．匿名加工とは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（値の削除や摂動化等）することであり，匿名加工されたデータから個人を識別する攻撃を再識別という．代表的なデータの加工技術として，k-匿名性 (k-anonymity)[1] や差分プライバシー (Differential Privacy)[4] などがある．

しかしながら，データのプライバシーリスクを正しく評価し，適切に匿名加工することは非常に困難である．例えば匿名加工の困難性として，有用性と安全性のトレードオフがあるからである．データは匿名加工することによって安全性を高めることができるが，それに伴い有用性は下がってしまう．説明のため，匿名加工の例を図 1.1 に示す．元データを X，X をそれぞれ異なる手法で匿名加工したものを A,B とする．元データである X は有用性が非常に高いが安全性は低く，個人が一意に特定される．一方，匿名加工データ B は 4 人全員のデータが同じになるように加工されているため，個人が特定される確率は低く安全性は高いが，X の情報を大きく加工しているため有用性は非常に低い．そこで，A のような，元データの有用性をある程度残し，かつ安全な匿名加工データが求められる（A がこの場合の最適な匿名加工であるとは限らない）．データのリスク評価指標・匿名加工手法の検討は，匿名加工・再識別コンテスト PWS CUP[5] で行われている．

匿名加工やプライバシーリスク評価に関する研究は国内外で盛んにおこなわれている．Klara らは，k-匿名性は一般性を欠いていることを主張し，これを一般化した n-混乱性 (n-confusion) を提案した [6]．Koot らは，カルバック・ライブラー距離 (Kullback-Leibler Distance) による近似を行い，デー

元データ X

氏名	性別	年齢	収入
A	男	25	300
B	男	28	350
C	女	30	400
D	女	32	450

匿名加工データ A

性別	年齢	収入
男	20代	325
男	20代	325
女	30代	425
女	30代	425

匿名加工データ B

性別	年齢	収入
-	20~40	300~500
-	20~40	300~500
-	20~40	300~500
-	20~40	300~500

図 1.1: 匿名加工データの例

タの匿名性を定量的に評価する指標を提案した [7]．Li らはデータマイニング後の攻撃者の知識が δ 以上増えないことを要求する指標 δ– プライバシー (δ-privacy) を提案した [8]．

早稲田らは攻撃者の背景知識の質や量に注目することにより，仮名化データのリスクを評価する指標を提案した [9]．南らは複数の集計表を用いた差分攻撃の解決策を論じ [10]，セル秘匿手法を提案している [11]．濱田らは匿名加工アルゴリズムが公開されることにより，データの危険度がどのくらい上がるのかを調査・分析した [12]．山田らは NP 困難であることが知られている k-匿名化問題を，ある条件のもと定数近似が可能であることを示した [13]．正木らは攻撃者の背景知識生成法と攻撃手法を提案することにより，軌跡情報の匿名性評価法を提案した [14]．また，特許庁は今後の進展が予想される技術テーマの一つとして匿名化技術を選定し，特許出願技術動向調査を実施している [15]．

1.3 既存研究と問題点

匿名加工における大きな課題として，攻撃者の想定がある．攻撃者はデータや個人についての背景知識（データの一部や統計情報等）を有しており，それをもとに匿名加工データから個人を再識別しようと試みることが想定される．

例えば最大知識攻撃者モデル（Maximum Knowledge Attacker Model）が Domingo-Ferrer らによって提案されており [16]，このモデルでは，元データと匿名加工データの全てを背景知識として持つ攻撃者が想定されている．元データをすべて持つ攻撃者が匿名加工データを攻撃する動機が疑問視されているが，このモデルはひとつの「最悪のケース（最強の攻撃者）」を想定するものであると著者らは主張している[1]．これにより，PWS CUP 等の様々な研究で攻撃者想定として採用されている．また最悪のケースのひとつとして，小栗らによって「最大能力攻撃者モデル」が提案されている [17]．このモデルでは，匿名加工アルゴリズム自体を背景知識として持つ攻撃者が想定されている．しかしながら，これらのような強すぎる攻撃者モデルは現実的ではなく，個人情報データや匿名加工データの危険性を過剰に評価してしまう恐れがある．例えば，攻撃者想定として最大知識攻撃者が採用された PWS CUP 2016[21] では，優勝チームの匿名加工データでさえ 22 %の購買履歴が正しく識別されてしまった．データの安全性を適切に評価するためには，より現実的な攻撃者モデルの想定が必要で

[1]Yet, since our maximum-knowledge attacker knows all original attribute values for all subjects, he has no reward to gain from the linkage. One might think of this attacker as being a purely malignant one (e.g. whose goal is to tarnish the data protector's reputation). Yet, the attacker's motivations are not important: we just want to model a worst-case attacker who has all the background information he can possibly use.

匿名加工データ C

仮名	性別	年齢	収入
1	男	25	300
2	男	28	350
3	女	30	400
4	女	32	450

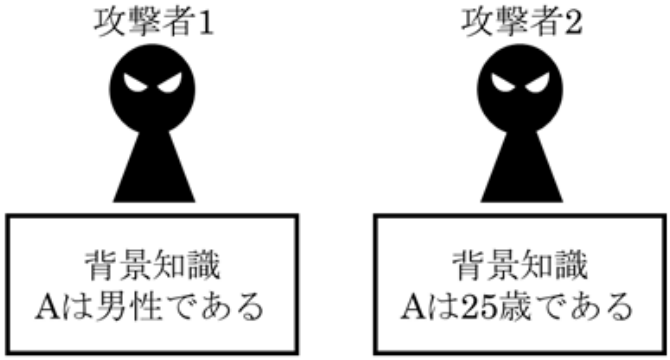

図 1.2: 攻撃者の有する背景知識の例

ある．

現実的な攻撃者モデルの例として，El Emam らによって提案されたモデルがある [18]．このモデルでは個人情報データが攻撃される 4 つのケース（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,「データが攻撃される確率」や「データ内に知り合いが含まれる確率」等を統計情報などから求めることにより，リスクを定式化している．しかしながら，このモデルではリスク計算に「データが攻撃される確率」や「データ侵害が発生する確率」といった主観的な値を必要としており，求めるのが困難である．これらの値はユースケースに大きく依存し，時や場合によって大きく変化する．

1.4 研究目的・研究方法

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．私は研究目的を達成するために，攻撃者の持つ背景知識に注目する．例えば図 1.2 の例を考えよう．匿名加工データ C は図 1.1 の元データ X を仮名化したものであり，攻撃者 1，2 はこのデータからユーザ A を特定しようとしている．攻撃者 1 は「A は男性である」という背景知識を有しているが，この背景知識からは A を一意に特定することはできない（ユーザ A ＝仮名 1 or 2）．一方,「A は 25 歳である」という背景知識を有している攻撃者 2 は A を一意に特定することができる（ユーザ A ＝仮名 1）．この場合，攻撃者 1 よりも攻撃種 2 の方が危険であるといえる．このように，攻撃者の危険度は背景知識の差（属性・量・質）に大きく依存する．本研究ではこの危険度の変化を研究する．

私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，元データのある属性から背景知識を得る攻撃者を想定する．背景知識を持つ攻撃者の識別確率の期待値（平均識別確率）をデータ中の属性の危険度とし，これを近似する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い，近似精度や計算コスト等の分析を行う．

1.5 本稿について

本稿の構成と，各章の概要は以下の通りである．また，表 1.1 に本稿におけるシンボル表を示す．

- 2 章 : 本稿にて用いるデータモデル，攻撃者モデル，リスクモデル（平均識別確率）を定義する．
- 3 章 : 購買履歴データに対する 10 タイプの具体的な攻撃者モデルを想定する．それらの平均識別確率を近似するモデルを提案し，実データを用いた評価実験を行う．
- 4 章 : 3 章のモデルを改善し，平均識別確率を近似する 3 つのモデル（平均モデル・最小コストモデル・サンプリングモデル）を提案する．また，複数の実データを用いて，これらのモデルの精度と計算コストについての評価を行う．
- 5 章 : 3 章のモデルの改善案として，エントロピーを用いたリスク評価指標を提案する．実際に収集した交通 IC カードデータを用いた評価実験を行う．
- 6 章 : 本稿のまとめを行う．

表 1.1: 本稿におけるシンボル表

シンボル	意味
T	履歴データ
T_{example}	履歴データの例
m	T のレコード数
n	T のユーザ数
X	T のある属性，5 章ではデータの用途
D_X	T の X の取りうる値の集合
ω_X	T における X のユニークな値の数，$\omega_X = \|D_X\|$
x	D_X の要素，背景知識
R_x	T で x を満たすレコード（行）インデックスの集合
U_x	T で x を満たすユーザの集合
T'	T を仮名化したデータ
$Pr(x)$	攻撃者が背景知識 x を得る確率
$Pr(\text{idf})$	個人が識別される確率
$Pr(\text{idf}\|x)$	攻撃者が背景知識 x から個人を識別する条件付き確率
$Pr(\text{idf}, x)$	$Pr(x)$ と $Pr(\text{idf}jx)$ の同時確率
α_x	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]
α_X	X における α_x の平均
u	T 内のあるユーザ
$Pr(\text{idf}, X)$	T の X のある値を背景知識として持つ攻撃者により個人が識別される確率の期待値，平均識別確率
T_1	購買履歴データ，Online Retail Data Set
$T_{1(id,day)}$	T_1 を顧客 ID，購買日について購買商品を整理したデータ
$T_{\text{example}(id,day)}$	T_{example} を顧客 ID，購買日について購買商品を整理したデータ
$Pr(\text{idf}, X, Y)$	属性 X と属性 Y についての背景知識を持つ攻撃者の平均識別確率
$R(X)$	平均識別確率の厳密解
$R_{\text{mean}}(X)$	平均モデルによって求められたリスク
$R_{\text{cost}}(X)$	最小コストモデルによって求められたリスク
$R_{\text{sample}}(X)$	サンプリングモデルによって求められたリスク
s	サンプリングサイズ
D'_X	D_X からランダムにサンプリングした s 個の要素の集合
$\alpha_{x'}$	D'_X についての α_x の平均
σ_s	s 個のサンプルの $R_{\text{sample}}(X)$ についての標準偏差
T_2	糖尿病患者の入院履歴データ，Diabetes Data Set
T_3	国税調査による世帯収入データ，Adult Data Set
M_{IC}	交通 IC カードから収集した顧客データ
T_{IC}	交通 IC カードから収集した履歴データ
E_S	ユーザごとの利用駅集計のデータ例
u_1, u_2, u_3	E_S 内のユーザ
s_1, s_2, s_3	E_S 内の駅名
$H(U)$	ユーザのエントロピー
$Pr(U = u_i)$	E_S 内での u_i の履歴の生起確率
$H(U\|S = s_i)$	U の条件付きエントロピー
$H(U\|S)$	ユーザの条件付きエントロピー
$I(U; S)$	1 つの駅利用履歴から得られる情報量の期待値，相互情報量
S	乗降用途
B	物販用途
C	チャージ用途
E_B	ユーザごとの購買集計のデータ例
$Pr(U\|X)$	用途 X の 1 履歴が与えられたときのユーザの平均識別確率

第2章　基礎定義

本章では，本稿にて用いるデータモデル，攻撃者モデル，リスクモデル（平均識別確率）を定義する．

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 2.1.1 履歴データを T とし，T のレコード数を m，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 2.1.1 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 *2.1* 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．また，T_{example} を仮名化した T'_{example} を表 *2.2* に示す．この場合，ユーザ *1*=仮名 *A*，ユーザ *2*=仮名 *B*，ユーザ *3*=仮名 *C* である．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 2.2.1 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．

表 2.1: 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

表 2.2: 仮名化された履歴データ T' の例 T'_{example}

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

定義 2.2.2 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 2.2.1,2.2.2 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 2.2.3 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 2.2.1 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 *2.3* 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 2.3.1 *(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 2.2.3 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 2.3.1 $X =$*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 2.3.2 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 2.3.2 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．

表 2.3: 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

第3章　背景知識の異なる10タイプの攻撃者と危険度

本章では，購買履歴データに対する 10 タイプの具体的な攻撃者モデルを想定する．それらの平均識別確率を近似するモデルを提案し，実データを用いた評価実験を行う．

3.1　Online Retail Data Set

本章にて用いるデータセットの説明・分析を行う．このデータは次章の評価実験でも用いる．UCI Machine Learning Repository[19] から公開されている Online Retail Data Set[20] は．英国の 1 年間の購買履歴のデータである．このデータは匿名加工・再識別コンテスト PWS CUP で用いられており，その際にコンテスト用に加工されている．本研究では PWS CUP 2016[21] で用いられた，400 人分の購買履歴データ T_1 を用いる．表 3.1 に T_1 の概要を示す．また，表 3.2 に T_1 の例を示し，表 3.3 に T_1 の統計量を示す．

本章で検討する攻撃者は，この履歴データについて，購買日，一日当たりの購買商品の種類数，一日当たりの購買商品の 3 種類の情報とその組み合わせの背景知識によって類型化されている．そこで，T_1 についてのこれらの属性の分布も以下に示し，表 3.4 にこれらの属性の統計量を示す．

図 3.1 に T_1 における購買日の出現頻度を示す．横軸の数値は月を示す．購買日の値域は 2010/12/1 から 2011/12/9 までの 373 日間であるが，そのうち購買があるのは 290 日であり，購買が存在しない日もある．図 3.2 に T_1 における購買種類数の出現頻度を示す．購買種類数とは 1 日の購買で同時に買われた商品の種類数である．例えば T_1 では，1 種類の購買が最も多く，71 回行われている．T_1 では 114 種類の種類数が生起する．図 3.3 に T_1 における購買商品の出現頻度を示す．横軸は頻度の順位を示す．例えば，T_1 で 2781 種類の商品が出現するが，最も多く購買されている商品は 1000 回以

表 3.1: 購買履歴データ T_1 の概要

属性	内容
顧客 ID	購買をした顧客の ID（5 桁数値）
伝票 ID	購買伝票の ID（6 桁数値）
購買日	購買した年月日（yyyy/mm/dd）
購買時	購買した時分（hh:mm）
購買商品	購買した商品の ID（数値，文字）
単価	購買した商品の単価 ($)
個数	商品を購買した個数

表 3.2: 購買履歴データ T_1 の例

顧客 ID	伝票 ID	購買日	購買時	購買商品	単価	個数
12583	536370	2010/12/1	8:45	22728	3.75	24
12583	536370	2010/12/1	8:45	22727	3.75	24
12431	536389	2010/12/1	10:03	22941	8.5	6
12431	536389	2010/12/1	10:03	21622	4.95	8
12431	536389	2010/12/1	10:03	21791	1.25	12
12838	536415	2010/12/1	11:57	22952	0.55	10
12567	537065	2010/12/5	11:57	22837	4.65	8
12567	537065	2010/12/5	11:57	22846	16.95	1
12748	537429	2010/12/6	15:54	84970S	0.85	12
12748	537429	2010/12/6	15:54	22549	1.45	8

表 3.3: 購買履歴データ T_1, $T_{example}$ の統計量

項目	Online Retail T	トイデータ $T_{example}$
レコード数 m	38087	10
顧客数 n	400	3
商品数 ω_{goods}	2871	4
購入日	2010/12/1-2011/12/9	2010/12/1-2010/12/3

上買われている.

3.2 背景知識の異なる１０タイプの攻撃者

購買履歴データ T_1 に対し，背景知識の異なる 10 タイプの攻撃者を想定する．購買履歴データ T_1 を顧客 ID・購買日について，購買商品を整理したデータ $T_{1(id,day)}$ を考える．攻撃者は，$T_{1(id,day)}$ からある顧客 u についての背景知識を得て，仮名化された T から u を識別しようと試みる場合を想定する．説明のために，表 2.1 の仮名化された簡易的な購買履歴データ T_{example} と，それを変換した表 3.5 の $T_{\text{example}(id,day)}$ を考えよう.

表 3.4: T_1 の 3 種類の情報の統計量

属性	購買日	購買種類数	購買商品
種類数 ω_X	290	114	2781
頻度平均値	5.4	13.75	13.7
頻度最大値	21	71	1012
頻度最頻値	5	1	1

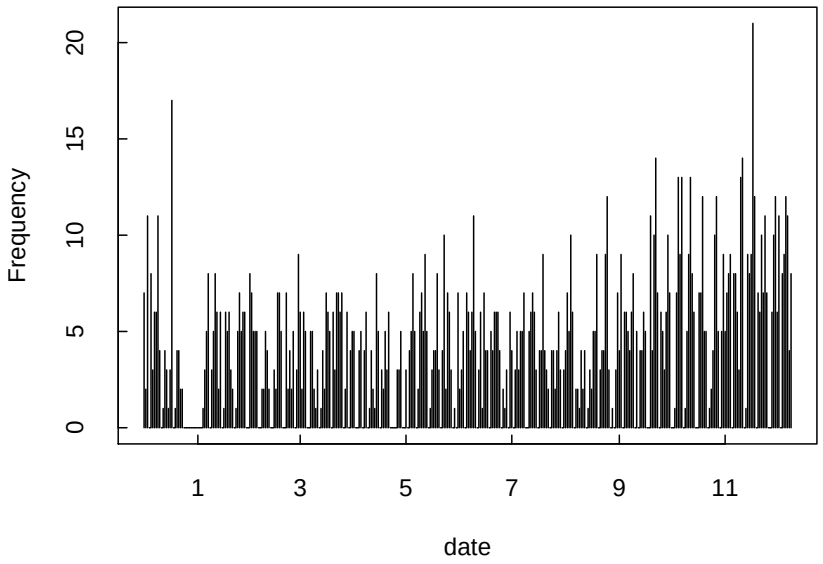

図 3.1: T_1 における出現頻度の日付分布

表 3.5: T_{example} のユーザ ID と購買日についての購買商品の表 $T_{\text{example}(id,day)}$

仮名 \ 購買日	2010/12/1	2010/12/2	2010/12/3
1	Bread, Book, Tea	Tea	
2	Bread		
3		Bread, Juice	Book, Tea, Juice

3.2.1 購買履歴データにおける背景知識

T_1 について，攻撃者が得る可能性のある背景知識について考える．T_1 の 7 属性は，大きく「誰が買ったか（顧客 ID，伝票 ID）」,「いつ買ったか（購買日，購買時）」,「何を買ったか（購買商品，単価，個数）」の 3 種類の情報に分類することができる．しかし 3 つのうち，攻撃者が「誰が買ったか」の情報を背景知識として得ることは考えにくいため，背景知識として得る可能性のあるのは「いつ買ったか」,「何を買ったか」についての情報であると考えられる．本研究では，これら 2 グループそれぞれの代表的な「購買日」,「購買商品」の 2 属性に加え，顧客が 1 日の購買で商品を「何種類買ったか」という情報に注目し，これらを攻撃者が得る背景知識として想定している．

「いつ買ったか」は知っている・知らないの 2 通り,「何種類買ったか」は知っている・知らないの 2 通り,「何を買ったか」は知らない・1 商品知っている・全商品知っているの 3 通りであるため，これらを組み合わせると 12 通りの攻撃者ができる．しかし,「何種類買ったか」を知らないのに「何を買ったか」を全て知っているのは矛盾するため，それらを除いて攻撃者のタイプは表 3.6 の 10 タイプになる．図 3.4 に各攻撃者の分類を示す．2 章ではデータの属性を攻撃者の背景知識としていたが，本章では「種類数」などの値も背景知識としていることに注意せよ．

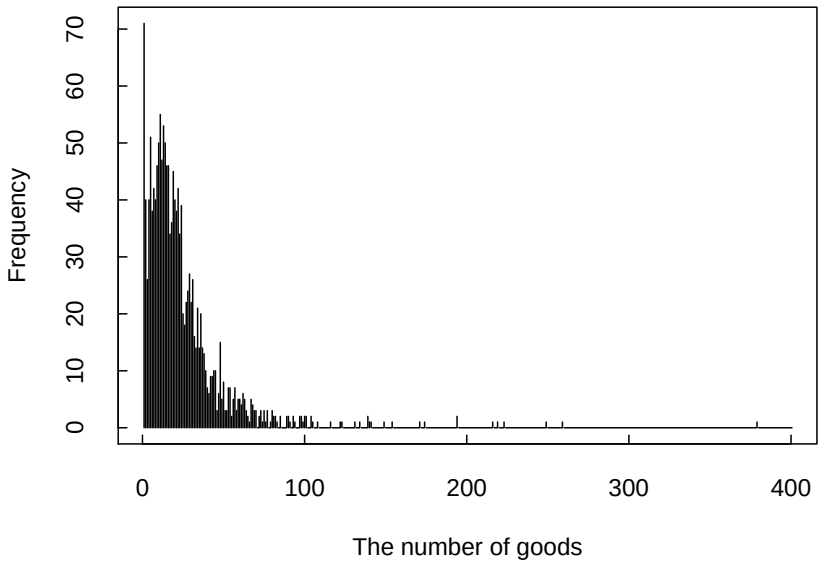

図 3.2: T_1 における一日当たりの購買商品種類数の頻度分布

3.2.2 10 タイプの攻撃者

攻撃者 0（無知識）

攻撃者 0 は背景知識を一切持っていない攻撃者である．T_{example} の顧客数は $n = 3$ であり，攻撃者 0 は n 人の顧客からランダムに u を識別するため，u を識別できる確率 $Pr(\text{idf})$ は $1/n = 1/3$ である．

攻撃者 1（1 商品）

攻撃者 1 は，u が購買した商品を 1 種類だけ知る攻撃者である．例えば攻撃者 1 が「u はお茶を買った」という背景知識を持っている場合，T_{example} でお茶を購買している顧客は仮名 1,3 の 2 人だけなので，このどちらかが u である．この場合，攻撃者 1 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{商品})$ は $1/2$ である．

攻撃者 2（何種類）

攻撃者 2 は，u が 1 日に何種類の商品を購買したかを知る攻撃者である．例えば攻撃者 2 が「u はある日，3 種類の商品を購買した」という背景知識を持っている場合，T_{example} で 1 日に 3 種類の商品を購買しているのは仮名 1,3 の 2 人だけなので，このどちらかが u である．この場合，攻撃者 2 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{種類数})$ は $1/2$ である．

攻撃者 3（何種類，1 商品）

攻撃者 3 は，u のある 1 日の購買商品種類数と，購買した商品を 1 種類だけ知る攻撃者である．例えば攻撃者 3 が「u はある日，3 種類の商品を購買し，その内 1 種類はお茶である」という背景知識

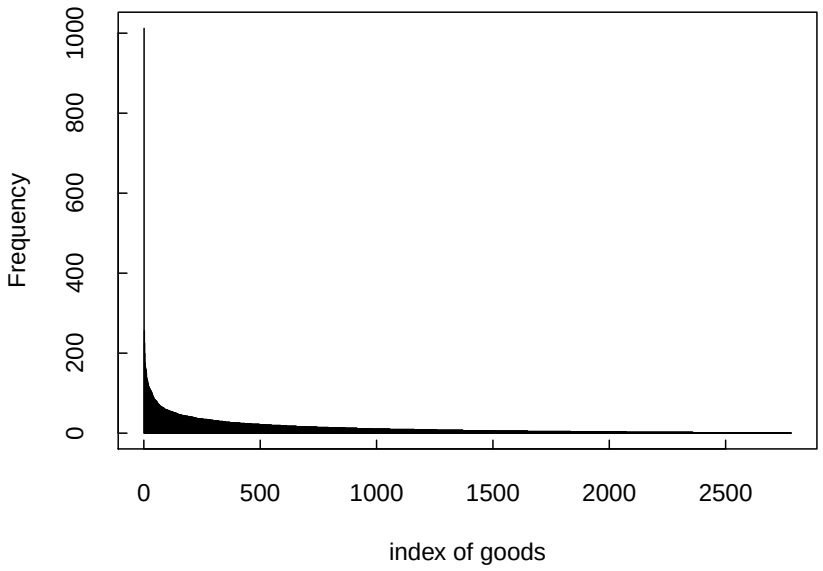

図 3.3: T_1 における購買商品の頻度分布

を持っている場合，T_{example} でこれに当てはまるのは仮名 1,3 の 2 人だけなので，このどちらかが u である．この場合，攻撃者 3 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{種類数}, \text{商品})$ は 1/2 である．

攻撃者 4（何種類，全商品）

攻撃者 4 は，u のある 1 日の購買商品種類数と，その日購買した商品を全て知る攻撃者である．例えば攻撃者 4 が「u はある日，3 種類の商品を購買し，その内容はパン，お茶，本である」という背景知識を持っている場合，T_{example} でこれに当てはまるのは仮名 1 だけなので，仮名 1 が u である．この場合，攻撃者 4 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{種類数}, \text{全商品})$ は 1 である．

攻撃者 5（いつ）

攻撃者 5 は，u がいつ購買したかを 1 日分だけ知る攻撃者である．例えば攻撃者 5 が「u は 2010/12/1 に購買をした」という背景知識を持っている場合，T_{example} で 2010/12/1 に購買をしているのは仮名 1,2 の 2 人だけなので，このどちらかが u である．この場合，攻撃者 5 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{購買日})$ は 1/2 である．

攻撃者 6（いつ，1 商品）

攻撃者 6 は，u がいつ購買したかを 1 日分と，その日購買した商品の 1 つを知る攻撃者である．例えば攻撃者 6 が「u は 2010/12/1 に本を買った」という背景知識を持っている場合，T_{example} でこれに当てはまるのは仮名 1 だけなので，これが u である．この場合，攻撃者 6 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{購買日}, \text{商品})$ は 1 である．

表 3.6: T_1 における 10 タイプの攻撃者

攻撃者	いつ	何種類	何を
0	×	×	×
1	×	×	1 商品
2	×	○	×
3	×	○	1 商品
4	×	○	全て
5	○	×	×
6	○	×	1 商品
7	○	○	×
8	○	○	1 商品
9	○	○	全て

攻撃者 7（いつ，何種類）

攻撃者 7 は，u がいつ購買したかを 1 日分と，その日に購買した種類数を知る攻撃者である．例えば攻撃者 7 が「u は 2010/12/1 に 1 種類の商品を買った」という背景知識を持っている場合，T_{example} でこれに当てはまるのは仮名 2 だけなので，仮名 2 が u である．この場合，攻撃者 7 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{購買日}, \text{種類数})$ は 1 である．

攻撃者 8（いつ，何種類，1 商品）

攻撃者 8 は，u がいつ購買したかを 1 日分と，その日に購買した種類数と，購買した商品を 1 種類だけ知る攻撃者である．例えば攻撃者 8 が「u は 2010/12/1 に 3 種類の商品を買い，そのうち 1 種類はパンである」という背景知識を持っている場合，T_{example} でこれに当てはまるのは仮名 1 だけなので，仮名 1 が u である．この場合，攻撃者 8 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{購買日}, \text{種類数}, \text{商品})$ は 1 である．

攻撃者 9（いつ，何種類，全商品）

攻撃者 9 は，u がいつ購買したかを 1 日分と，その日に購買した種類数と，購買した商品を全て知る攻撃者である．例えば攻撃者 9 が「u は 2010/12/1 に 3 種類の商品を買い，その内容はパン，本，お茶である」という背景知識を持っている場合，T_{example} でこれに当てはまるのは仮名 1 だけなので，仮名 1 が u である．この場合，攻撃者 9 が背景知識から u を識別できる確率 $Pr(\text{idf}|\text{購買日}, \text{種類数}, \text{全商品})$ は 1 である．

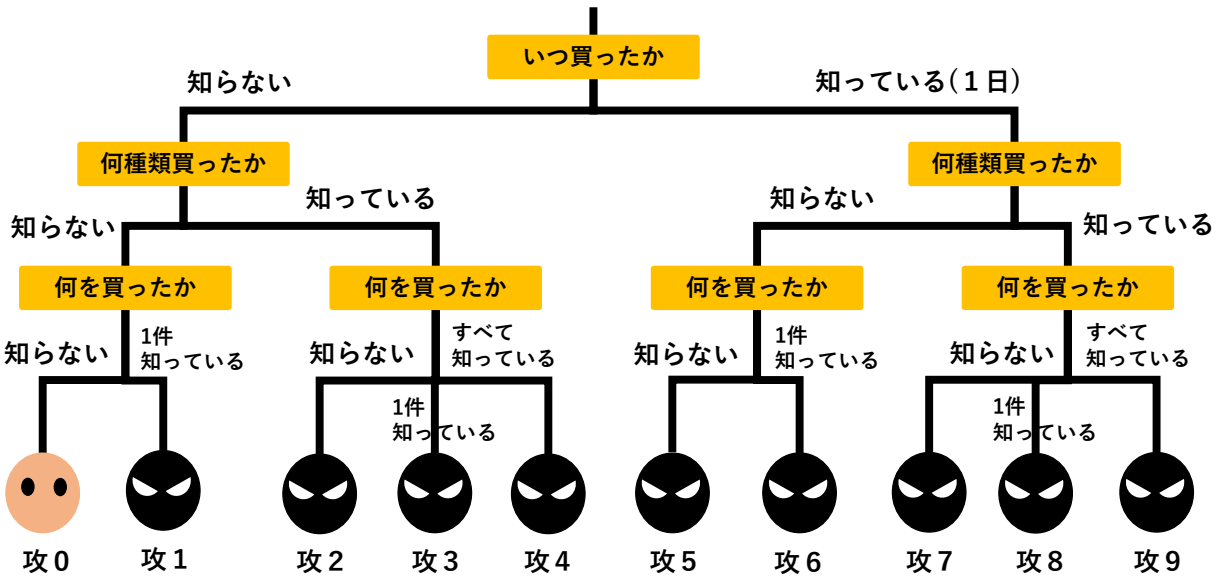

図 3.4: T_1 における 10 タイプの攻撃者の分類

3.3 攻撃者の危険度の評価

10 タイプの攻撃者の危険度を，2 章にて定義した平均識別確率を用いて評価する．しかしながら，ビッグデータに対して平均識別確率を求めるのは困難であるので，これを近似する方法を提案する．

3.3.1 平均識別確率の近似

仮定 3.3.1 任意の背景知識 x について，$|R_x| = |U_x|$ である．すなわち，$\alpha_x = 1$.

例 3.3.1 T_{example} の X=*Date* の場合，x=*2010/12/1* のとき $|R_x| = 4$, $|U_x| = 2$ であるため，仮定 *3.3.1* は成立しない．

「いつ買ったか」,「何種類買ったか」,「何を買ったか」のうち，1 種類のみの背景知識 x を持つ攻撃者の平均識別確率は次の定理で与えられる．4 章ではこの定理によって求められるリスクを最小コストモデルとしていることに注意せよ．

定理 3.3.1 仮定 *3.3.1* のもと，購買履歴データ T_1 について，単一の背景知識 X を持つ攻撃者の平均識別確率 $Pr(\text{idf}, X)$ は，

$$Pr(\text{idf}, X) = \frac{\omega_X}{m}$$

である．

(Proof) T_1 の属性 X についての背景知識を持つ攻撃者の平均識別確率は

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

である．このとき，$|U_x| = |R_x|$ と仮定すると，

$$Pr(\mathrm{idf}, X) = \sum_{x \in D_X} \frac{1}{m} = \frac{\omega_X}{m}$$

となり，定理 3.3.1 が成り立つ． (Q.E.D)

例 3.3.2 T_{example} を用いて定理 *4.1* の説明を行う．$X =$*Date* とする．例えば $x =$「*2010/12/1* に買った」とすると，$R_x = \{1, 2, 3, 4\}$, $U_x = \{1, 2\}$ である．また，$m = 10$，$\omega_X = 3$ であるため，x を背景知識として持つ攻撃者の平均識別確率は $1/|U_x| = 1/2$ であり，その生起確率は $|R_x|/m = 4/10$ である．よって，*Date* を背景知識として持つ攻撃者の平均識別確率は

$$Pr(\mathrm{idf}, X) = \sum_{x \in D_X} \frac{|R_x|}{m} \frac{1}{|U_x|} = \frac{4 + 3 + 6}{20} = 0.65$$

となる．このとき，仮定 *3.3.1* をおくと，

$$Pr(\mathrm{idf}, X) = \sum_{x \in D_X} \frac{1}{m} = \frac{\omega_X}{m} = 0.3$$

となる．

定理 3.3.2 仮定 *3.3.1* のもと，購買履歴データ T_1 について，属性 X についての背景知識 x と，独立な属性 Y についての背景知識 y を同時に持つ攻撃者の平均識別確率 $Pr(\mathrm{idf}, X, Y)$ は，

$$Pr(\mathrm{idf}, X, Y) = \frac{\omega_X \omega_Y}{m}$$

である．

(Proof) x を D_X の要素とし，y を D_Y の要素とする．T_1 で x が生起しているレコードの集合を R_x，T で x に当てはまるユーザの集合を U_x としたとき，背景知識 x, y が同時に発生する確率は独立性を仮定すると，$Pr(x, y) = Pr(x)Pr(y) = \frac{|R_x|}{m}\frac{|R_y|}{m}$ である．また，仮定 3.3.1 をおくと，それらの背景知識に当てはまる人数は $m\frac{|R_x|}{m}\frac{|R_y|}{m}$ であるため，リスクは $\frac{1}{m\frac{|R_x|}{m}\frac{|R_y|}{m}}$ となる．よって，x, y を同時に持つ攻撃者の平均識別確率は

$$\begin{aligned} Pr(\mathrm{idf}, X, Y) &= \sum_{x \in D_X} \sum_{y \in D_Y} \frac{\frac{|R_x|}{m}\frac{|R_y|}{m}}{m\frac{|R_x|}{m}\frac{|R_y|}{m}} \\ &= \sum_{x \in D_X} \sum_{y \in D_Y} \frac{1}{m} \\ &= \frac{\omega_X \omega_Y}{m} \end{aligned}$$

となり，定理 3.3.2 が成り立つ． (Q.E.D)

例 3.3.3 T_{example} を用いて定理 *4.2* の説明を行う．$X =$「いつ買ったか」とし，$Y =$「何種類買ったか」とする．例えば $x =$「*2010/12/1* に買った」，$y =$「*1* 種類買った」とすると，$R_x = \{1, 2, 3, 4\}$，$U_x = \{1, 2\}$，$R_y = \{1, 2\}$，$U_x = \{4, 5\}$ である．また，$m = 10$，$\omega_X = 3$，$\omega_Y = 3$ であるため，x と y を背景知識として持つ攻撃者の識別確率は仮定 *3.3.1* をおき，独立性を仮定すると $\frac{1}{m\frac{|R_x|}{m}\frac{|R_y|}{m}} = \frac{10}{8}$ であり，その生起確率は $\frac{|R_x|}{m}\frac{|R_y|}{m} = \frac{8}{100}$ である．よって,「いつ買ったか」と「何種類買ったか」を背景知識として持つ攻撃者の平均識別確率は $Pr(\mathrm{idf}, X, Y) = \frac{\omega_X \omega_Y}{m} = 0.9$ となる．

表 3.7: T_1 に対する攻撃者 0〜9 の危険度

攻撃者	実測値	近似値
0	0.0025	0.0025
1	0.0965	0.0730
2	0.0807	0.0030
3	0.7974	8.3239
4	0.9788	4.5436
5	0.1851	0.0076
6	0.8945	21.1749
7	0.9400	0.8680
8	0.9750	2413.9433
9	0.9994	1317.6433

定理 3.3.3 仮定 *3.3.1* のもと，購買履歴データ T_1 について，独立に発生する複数の属性 $X_1, X_2, \cdots, X_k$ についての背景知識を持つ攻撃者の平均識別確率 $Pr(\mathrm{idf}, X_1, X_2, \cdots, X_k)$ は，

$$Pr(\mathrm{idf}, X_1, X_2, \cdots . X_k) = \frac{\omega_{X_1}\omega_{X_2}\cdots\omega_{X_k}}{m}$$

である.（証明略）

3.3.2 Online Retail Data Set についての攻撃者の危険度の実測値

表 3.7 に購買履歴データ T_1 に対する各攻撃者の危険度を示す．実測値は平均識別確率 $Pr(\mathrm{idf}, X)$ によって求めた攻撃者の危険度であり，近似値は前節の定理を用いて求めた値である．T_1 の場合，表 3.4 より $\omega_{day} = 290$，$\omega_{num} = 114$，$\omega_{goods} = 2781$ である．

攻撃者 1,2,5 は順に「何種類買ったか」,「何を買ったか（1 商品）」,「いつ買ったか」を背景知識として知る攻撃者だが，その中では攻撃者 5 の危険度が最も高く，0.1851 であった．また，攻撃者の危険度は複数の種類の背景知識を持つことで大きく上昇した．

3.3.3 考察

表 3.8 に平均識別確率の実測値と近似値に基づく攻撃者のランクを示す．どちらの場合も最も危険度が低いのは攻撃者 0（何も知らない）であるが，最も危険度が高い攻撃者は異なり，近似値の場合は攻撃者 8（いつ，何種類，商品 1 品を知っている），実測値の場合は攻撃者 9（いつ，何種類，商品全てを知っている）となった．しかし，攻撃者 8 の背景知識量は明らかに攻撃者 9 より少なく，危険度も攻撃者 9 の方が高くなるのが自然である．

また，表 3.7 より，平均識別確率の実測値と近似値は誤差が非常に大きく，近似値の値が 1 を超えてしまうものもある．その理由は，不自然な仮定をおいていることだと考えられる．

近似値では 2 つの大きな仮定を置いている.「背景知識の生起する回数 $|R_x|$ と当てはまる人数 $|U_x|$ が等しいこと（$|U_x| = |R_x|$）」と「独立性（$Pr(x, y) = Pr(x)Pr(y)$）」である．図 3.5，3.6，3.7 に，

表 3.8: 実測値と近似値についての攻撃者の危険度順位

順位	攻撃者（実測値）	攻撃者（近似値）
1	**9**	**8**
2	4	**9**
3	**8**	6
4	7	3
5	6	4
6	3	7
7	5	1
8	1	5
9	2	2
10	0	0

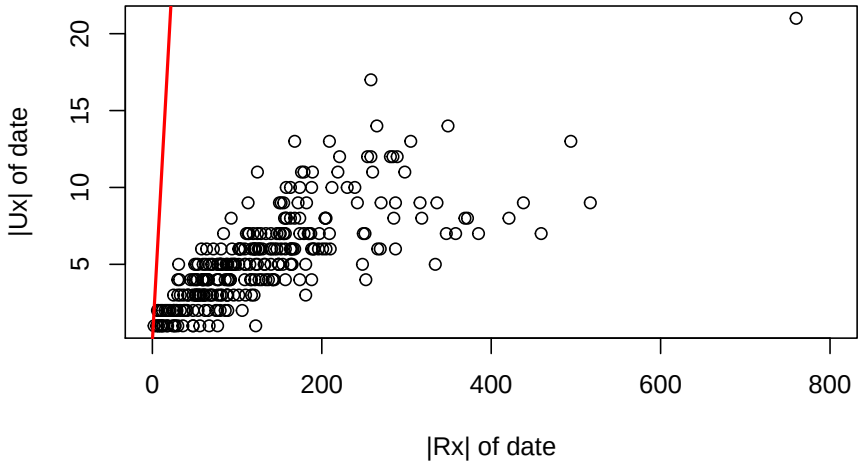

図 3.5: 購買日についての $|R_x|$ と $|U_x|$ の散布図

それぞれ「いつ買ったか」,「何種類買ったか」,「何を買ったか」についての，$|R_x|$ と $|U_x|$ の散布図を示す．赤い直線は $|U_x| = |R_x|$ を表しており，ほとんどの点 (x) について $|U_x| = |R_x|$ を満たしていないことがわかる．

また，独立性の仮定にも大きな問題がある．独立性を仮定すると，背景知識「x かつ y」の生起確率は $p_x p_y$ となり，これは 0 にならない．しかし実際のデータでは背景知識 $x \in D_X$ と $y \in D_Y$ の組み合わせの中には生起しないものも非常に多い．例えば T の購買日と購買種類数に注目した $T_{(day,num)}$ を考えると，$\omega_{day} = 290$，$\omega_{num} = 114$ より，$\omega_{day}\omega_{num} = 33060$ となるが，このうち背景知識として生起するのは 1473 種類のみ（0.46%）であった．このことより，独立性の仮定は不自然であると考えられる．

本章では，ある顧客の 3 種類の属性についての情報の 1 日分を背景知識として持つ弱い攻撃者を想定している．しかし現実の場合は，攻撃者は複数日・複数人分の情報を背景知識として持つことが想定できる．また，攻撃者が「何を買ったか」についての情報を「知らない」,「1 商品知っている」,「全て知っている」の 3 つの場合に分類しているが,「k 商品知っている」ときのリスクは想定できて

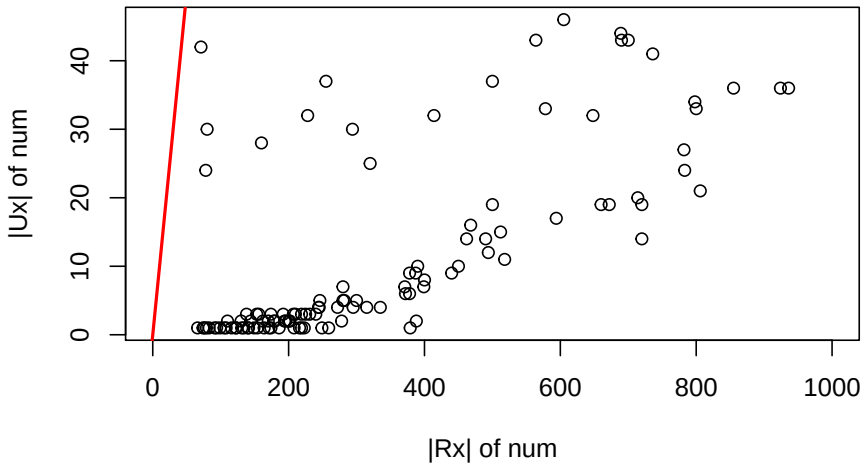

図 3.6: 購買種類数についての $|R_x|$ と $|U_x|$ の散布図

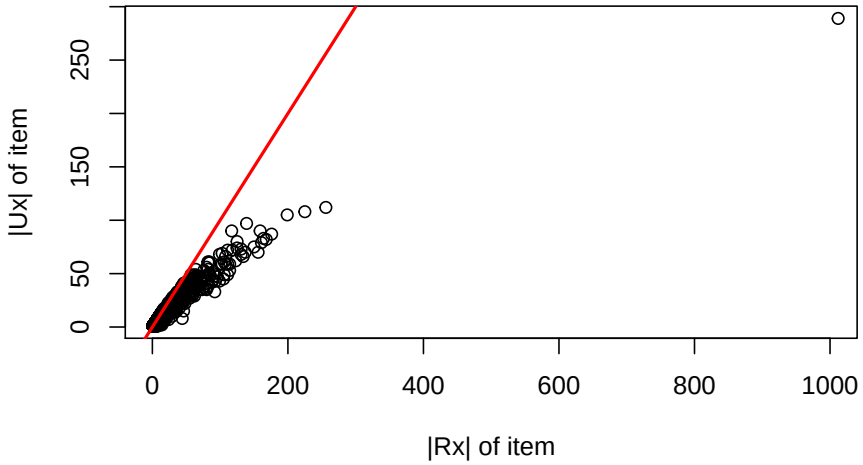

図 3.7: 購買商品についての $|R_x|$ と $|U_x|$ の散布図

いない．

4 章ではこのモデルを改善したモデルを提案し，様々なデータセットに適応する．また 5 章では，複数の背景知識を持つ攻撃者のリスクをエントロピーを用いて評価する．

第4章　平均識別確率を近似するリスク評価モデル

3章では，購買履歴データ T_1 に対する具体的な10タイプの攻撃者を想定し，これらの平均識別確率を近似した．しかしながら，近似のためにおいた仮定が不自然であり，近似精度が悪かった．本章ではこれを改善するために新たに3つの近似モデルを提案する．また3章ではリスク評価の対象を購買履歴データに限定し，具体的な攻撃者を想定していたが，本章では様々なデータに対する攻撃者を想定し，T_1 を含む3つのデータセットを用いた評価実験を行う．

4.1　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2.3.1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

4.1.1　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，$R(X) = Pr(\mathrm{idf}, X)$ 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．

4.1.2　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 4.1.1 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．

興味深いことに，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 4.1.1 $R_{mean}(X)$ を定義 *4.1.1* 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，$R_{mean}(X) = Pr(\mathrm{idf}, X)$ である．

(Proof) 定義 4.1.1,2.2.3 より，

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

であり，定理 4.1.1 を得る． (Q.E.D)

例 4.1.1 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．

4.1.3 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 $\alpha_x = 1$（仮定 3.3.1 を満たす）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 4.1.2 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 4.1.2 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 4.1.2 最小コストモデルの誤差率は $|\frac{1}{\alpha_X} - 1|$ である．

(Proof) 定理 4.1.1 により $Pr(\mathrm{idf}, X) = \alpha_X \omega_X / m$ を用いると，$R_{cost}(X)$ の厳密解に対する誤差率は，

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

となるため，定理 4.1.2 を得る． (Q.E.D)

定義 2.2.1 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．

4.1.4 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．以下のように定義を行う．

定義 4.1.3 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 4.1.3 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 4.1.3 サンプリングモデルの誤差率の最大値は

$$\frac{\sigma_s m}{\sqrt{|s|} \omega_X \alpha_X} \tag{4.1}$$

である．

(Proof) 定理 4.1.1 より，$Pr(\mathrm{idf}, X) = \alpha_X \omega_X / m$ である．このとき，90%信頼区間を仮定すると，$Pr(\mathrm{idf}, X)$ との絶対誤差は $|R_{sample}(X) - Pr(\mathrm{idf}, X)| < Var[Pr(\mathrm{idf}, X)] = \sigma_s / \sqrt{s}$ となる．よって，$R_{sample}(X)$ と厳密解の相対誤差率は

$$\begin{aligned} &= \frac{|R_{sample}(X) - Pr(\mathrm{idf}, X)|}{Pr(\quad \mathrm{idf}, X)} \\ &= \frac{\frac{1}{\sqrt{s}} \sigma_s}{\frac{\alpha_X \omega_X}{m}} = \frac{\sigma_s m}{\sqrt{|s|} \omega_X \alpha_X} \end{aligned}$$

表 4.1: 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

表 4.2: 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

となるため，定理 4.1.3 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 4.1 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．

4.2 評価実験

4.2.1 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[19] 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 [20]
2. T_2: Diabetes Data Set [22]
3. T_3: Adult Data Set [23]

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 4.2 に示す．

4.2.2 データセットの分析

表 4.3 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 4.1〜4.5 に示し，

表 4.3: 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)

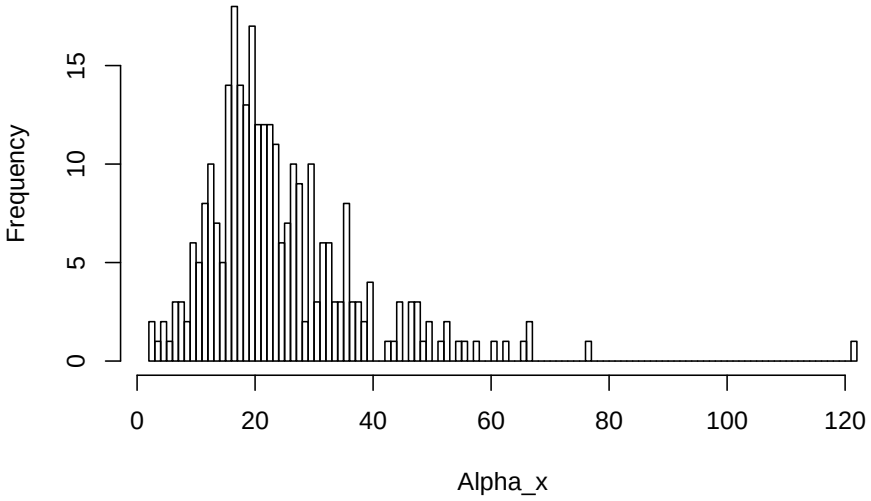

図 4.1: T_1 の X =Date 属性についての α_x の分布

各属性についての α_X と ω_X を表 4.4 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 4.6,4.7 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 4.5 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 4.6 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 4.4 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 4.8 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.

表 4.4: 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	$1.54 \cdot 10^{-4}$	$4.79 \cdot 10^{-4}$

表 4.5: 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

4.2.3 平均識別確率によるリスク評価結果

T_1, T_2, T_3 の各属性の危険度を厳密解によって求める．$R(X)$ と $Pr(\mathrm{idf}|x)$ の標準偏差 δ を表 4.4 に示す．例えば，T_1 の Date 属性の平均識別確率は 0.186 であり，$\delta = 0.140$ であった．T_1 で最も危険であると評価されたのは Time 属性であり，平均識別確率は 0.322 であった．平均識別確率によるリスク評価は，データを匿名加工する際の指標になり，この場合はまず Time 属性を，丸め込み (8:45→8:00) や摂動化 (8:45→8:42) や削除 (8:00→NA) といった手法で加工する必要がある．Time 属性や Date 属性のような時間に関連する属性は，匿名加工の際に削除されることが多い（例：[18] の 5 章）．

T_2, T_3 は ω_X, α_X がともに小さいため，平均識別確率も小さくなった．T_2 では Time 属性が，T_3 では Age 属性が最も危険であると評価された．T_2, T_3 は平均レコード数の平均 α_X が 1.0 に近く，このようなデータにおいては平均識別確率 $Pr(\mathrm{idf}, X)$ と値の種類数 ω_X の順位がほぼ等しくなる．一方，T_1 のような α_X が大きいデータのリスク評価には注意を払う必要がある．このように，提案したリスクモデルは様々なデータに適用できる．

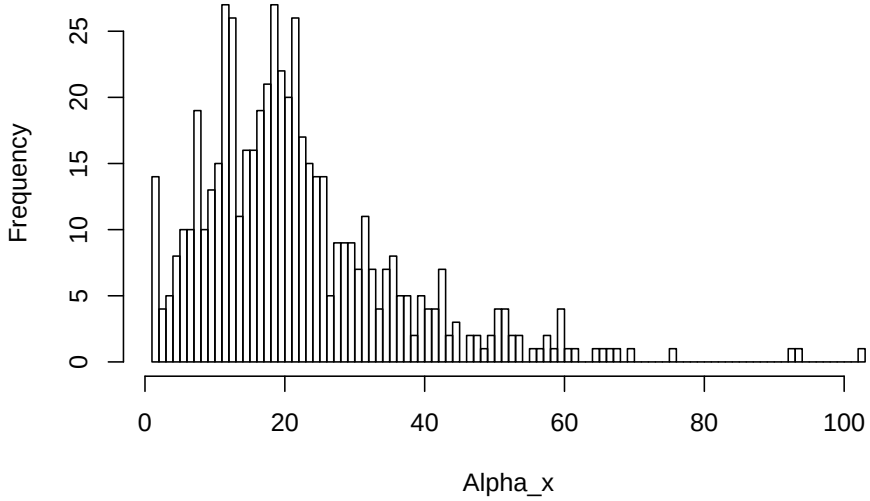

図 4.2: T_1 の X =Time 属性についての α_x の分布

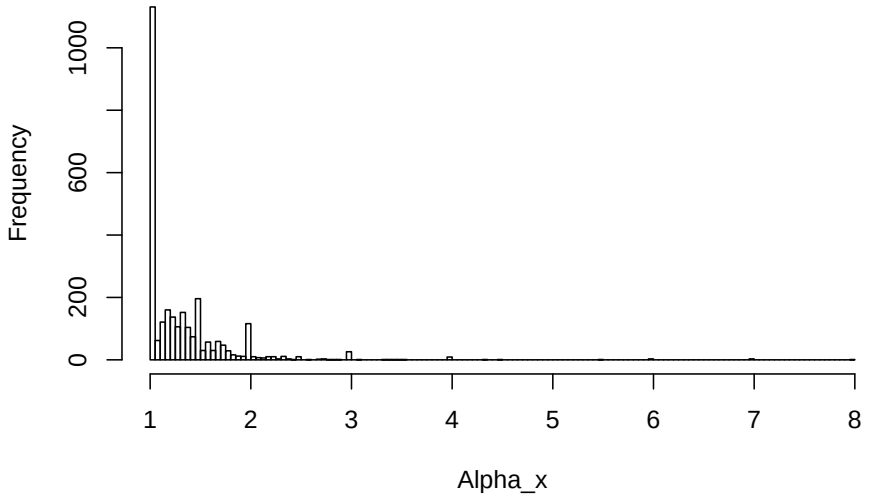

図 4.3: T_1 の X =Goods 属性についての α_x の分布

4.2.4 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 4.7 に各モデルの評価値を示す．定理 4.1.1 により，平均モデルによる評価値 $R_{mean}(X)$ は表 4.4 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，$s = 10$ のときの $90\%(\mu \pm \delta)$ 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 4.8 に各モデルのコストと誤差の値を示し，図 4.10 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

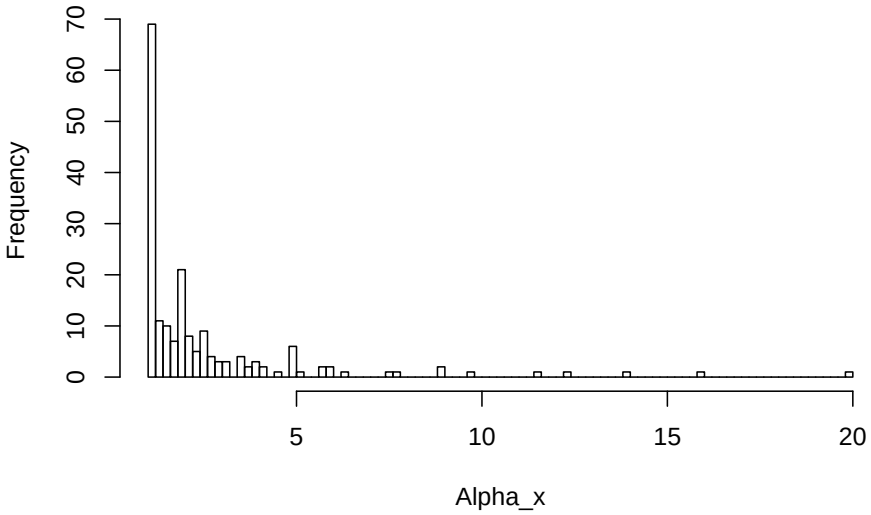

図 4.4: T_1 の X =Price 属性についての α_x の分布

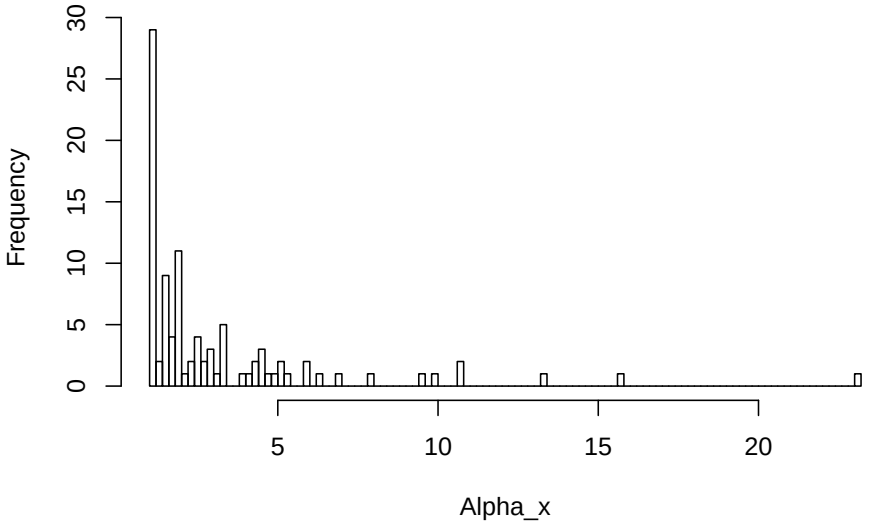

図 4.5: T_1 の X =Number 属性についての α_x の分布

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 4.11 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 4.9 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．

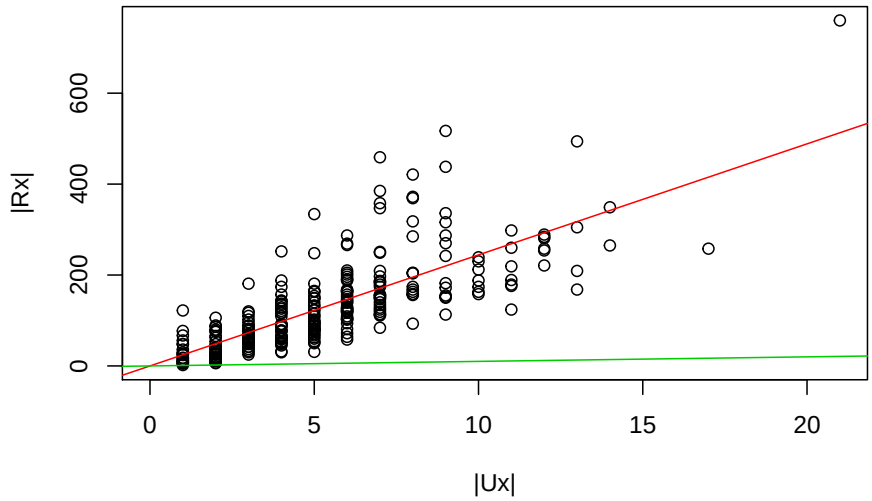

図 4.6: T_1 の X =Date 属性についての散布図

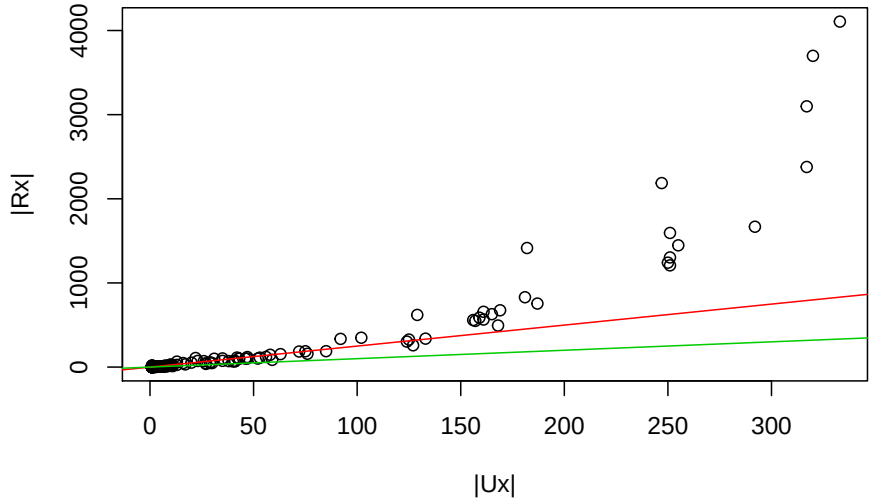

図 4.7: T_1 の X =Price 属性についての散布図

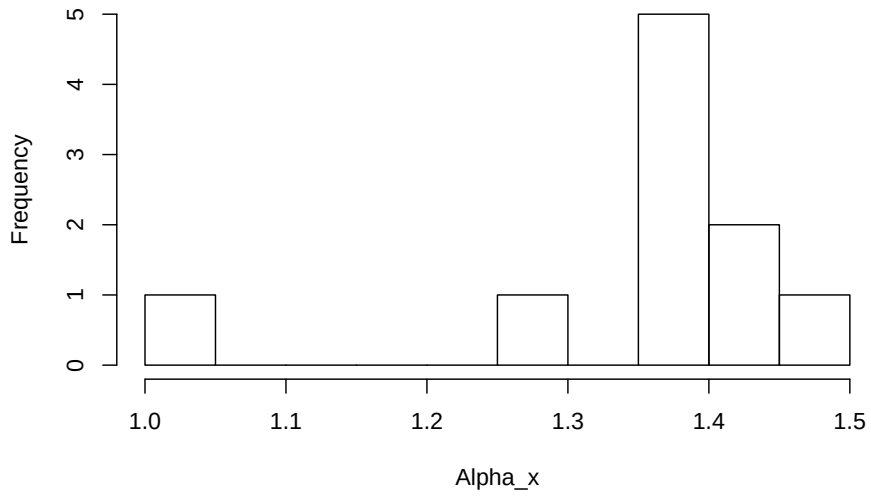

図 4.8: T_2 の X =Age 属性についての α_x の分布

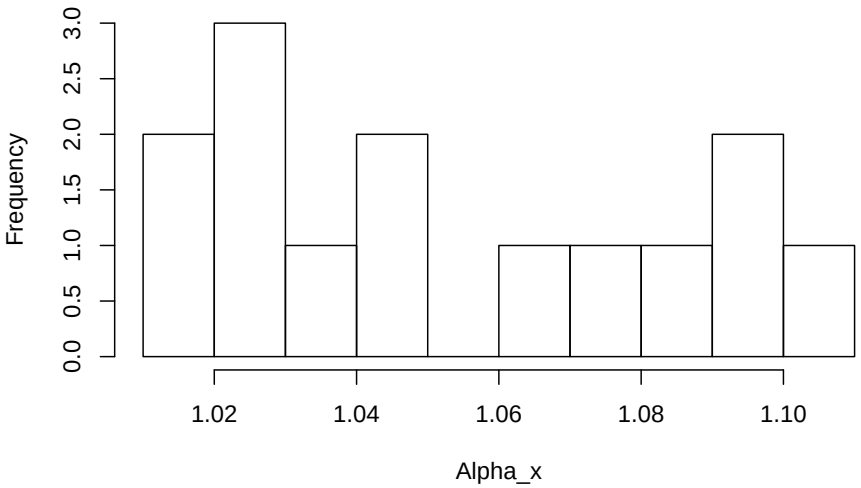

図 4.9: T_2 の X =Time 属性についての α_x の分布

表 4.6: 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

表 4.7: 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 4.8: 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

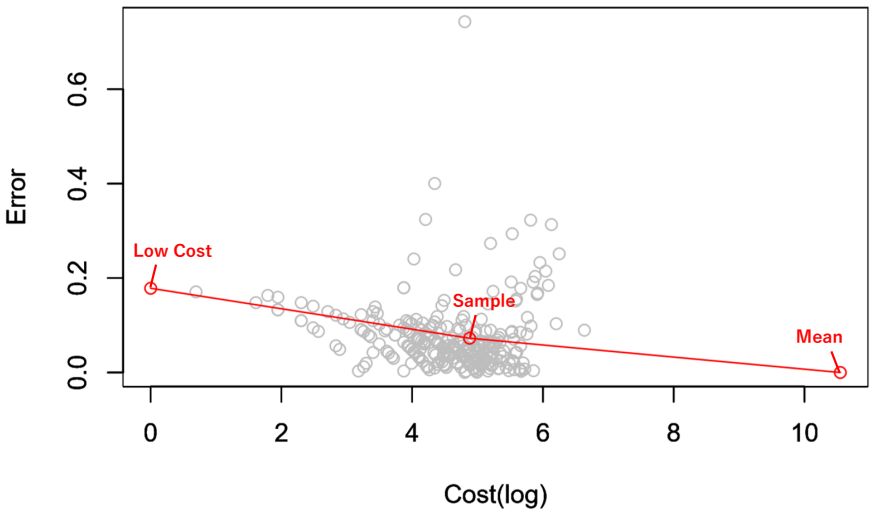

図 4.10: 各モデルのコストと誤差の散布図

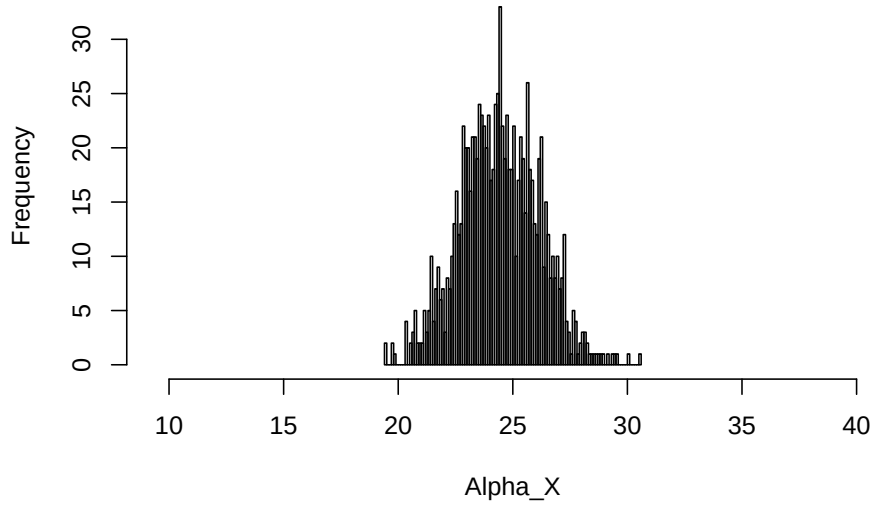

図 4.11: T_1 の Date 属性から 50 グループをサンプリングしたときの α_X の分布

表 4.9: 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

第5章　エントロピーを用いたリスク評価手法の提案

4 章にて，履歴 T のある単一の属性 X についての背景知識 x を持つ攻撃者の平均識別確率を近似するモデルを提案し，評価実験を行った．しかしながら，3 章における攻撃者 6 のような，複数の属性についての背景知識を持つ攻撃者の危険度についてはいまだ精度よく近似できていない．本章ではこれを解決するために，エントロピーを用いたリスク評価指標を提案し，交通 IC カードデータを用いて評価実験を行う．

5.1　交通 IC カード履歴データの取得

本研究のために，明治大学総合数理学部に所属する 31 人から同意を取り，交通 IC カードから顧客データ M_{IC} と利用履歴データ T_{IC} 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 [24]」を使用した．一人あたりから収集できる履歴は最大 19 件である．表 5.1 にアプリケーションで取得できる乗降履歴データ T_{IC} の例を示す．なお，このデータは著者の卒業論文である [25] にて用いたものと同一である．

表 5.2 に取得した本データの概要を示す．顧客データ M_{IC}(マスターデータ) は 31 レコード 6 属性のデータであり，履歴データ T_{IC}(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 $(m = 584, n = 31)$．表 5.3 に顧客データの例を示す．表 2.1 に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 5.1 の履歴をデータ化したものが表 5.4 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_{IC} 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_{IC} に定期券の範囲を加えた．

5.2　識別リスク分析

5.2.1　リスクの考え方

データのリスク評価には様々な方法があるが，本節ではデータのエントロピー [bit/symbol] に注目してリスク評価を行う．表 5.5 のデータ例 E_S を用いて考え方を説明する．E_S は 3 人のユーザの計

表 5.1: 取得できる履歴データの例

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 5.2: 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	ℓ 584 件	顧客 ID	2 桁数値
			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

19 回の駅利用履歴データについての集計表である．u_1, u_2, u_3 はユーザ，s_1, s_2, s_3 は駅名である．例えば u_1 は s_1 を 2 回，s_2 を 1 回利用している．

はじめに,「駅利用履歴が完全に不明である場合」のユーザのエントロピー $H(U)$ は，$Pr(U = u_i)$ をデータ E_S 中で u_i の履歴の生起確率，n をユーザー数としたとき，

$$H(U) = -\sum_{i=1}^{n} Pr(U = u_i) \log_2 Pr(U = u_i)$$

で与えられる．E_S の場合，全 19 履歴のうち，u_1 のものは 3 回であるため，$Pr(U = u_1) = 3/19$,$Pr(U = u_2) = 8/19$,$Pr(U = u_3) = 8/19$ である．よって，$H(U) = 1.47$[bit/履歴] となる．

次に,「駅利用履歴 S が与えられた場合」のユーザの条件付きエントロピー $H(U|S)$ を考える．$Pr(S = s_i)$ を E_S 中で駅 s_i の履歴の生起確率，ω_{station} を駅の種類数とすると，U の条件付きエントロピーは

$$H(U|S = s_i) = -\sum_{j=1}^{n} Pr(U = u_j|S = s_i) \log_2 Pr(U = u_j|S = s_i)$$

表 5.3: 顧客データ M_{IC} の例

顧客 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

表 5.4: 履歴データ T_{IC} 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

で与えられ，ユーザのエントロピーは

$$H(U|S) = \sum_{i=1}^{\omega_{\mathrm{station}}} Pr(S = s_i) H(U|S = s_i)$$

で与えられる．E_S の場合，

$$H(U|S) = \sum_{i=1}^{3} Pr(S = s_i) H(U|S = s_i) = \frac{10}{19} \cdot 1.52 + \frac{5}{19} \cdot 0.72 = 0.99$$

である．全 19 履歴のうち，s_1 のものは 10 回，s_2 のものは 5 回，s_3 のものは 4 回であるため，$Pr(s_1) = 10/19$，$Pr(s_2) = 5/19$，$Pr(s_3) = 4/19$ であり，$H(u|s_1) = 1.52$，$H(u|s_2) = 0.72$，$H(u|s_3) = 0$ である．よって，$H(U|S) = 0.99$ となる．

最後に，相互情報量 $I(U;S)$ を求める．相互情報量とは 1 つの駅利用履歴から得られる情報量の期待値であり，$I(U;S) = H(U) - H(U|S)$ で与えられる．E_S の場合，$I(U;S) = 0.48$ である．

$H(U), H(U|S), I(U;S)$ の意味を考える．例えば駅利用履歴が完全に不明である場合，$H(U) = 1.47$ であり，u_1, u_2, u_3 の中のあるユーザを特定できる平均識別確率 $Pr(\mathrm{idf})$ は $1/2^{H(U)} = 0.36$ である．しかし，1 つの駅利用履歴が判明した場合，例えば，s_3 が分かると一意に u_2 であることが特定されるが，s_2 ならば u_1 か u_3 らしいことしか分からない．平均すると $H(U|S) = 0.99$ になり，その平均識別確率 $Pr(\mathrm{idf}|S)$ は $1/2^{H(U|S)} = 0.5$ である．このとき 1 つの駅利用履歴から得た情報量は $I(U;S) = 0.48\mathrm{bit}$ であるため，

$$H(U) = 1.47 < 1.92 = 4I(U;S)$$

より，4 つの駅利用履歴が判明した場合，u_1, u_2, u_3 の中の全てのユーザを特定できる確率はほぼ 1 になる．

駅利用履歴が不明である場合のユーザのエントロピーは $H(U) = -\sum_{i=1}^{n} Pr(U_i) \log_2 Pr(U_i)$ で求められる．$Pr(U_i)$ はデータ中で U_i の履歴が登場する確率であり，n はユーザー数である．また，駅利用履歴が 1 つ判明した場合の平均エントロピーは $H(U|S) = H(U) - \sum_{i=1}^{\omega_X} Pr(S_i) * H(S_i)$ で求められる．この場合，ω_X は駅の種類数である．相互情報量は $I(U;S) = H(U) - H(U|S)$ で求められる．$I(U;S)$ はそのデータの 1 つの履歴から得られる情報量の大きさの期待値を示している．これらの値をデータのリスク評価に用いる．

表 5.5: ユーザごとの利用駅集計のデータ例 E_S

ユーザ \ 駅	s_1	s_2	s_3	計	$Pr(U=u_i)$
u_1	2	1	0	3	3/19
u_2	4	0	4	8	8/19
u_3	4	4	0	8	8/19
$H(U\|S=s_i)$	1.52	0.72	0		
$Pr(S=s_i)$	10/19	5/19	4/19		

E_S の場合，$H(U)=1.47$,$H(U|S)=0.99$,$I(U;S)=0.48$ である．$H(S_i)$ の値が大きいほどその属性から得られる情報量は少ない．例えば不明なユーザ U^* の交通 IC カードから S_1 の利用履歴を得た場合，S_1 は全ユーザが利用しているため，U^* を一意に特定することはできない（$H(S_1)=1.52$，情報量が少ない）．しかし，得た履歴が S_3 のものであった場合，S_3 は U_2 しか利用していないため，$U^*=U_2$ と一意に特定することが可能である（$H(S_3)=0$，情報量が大きい）．

5.2.2 多様なエントロピー

取得した履歴データの，用途「交通」,「物販」,「チャージ」についてまとめた集計表を順に D_S, D_B, D_C とする．ユーザ (U) の数は 31 人，利用駅 (S) の種類は 138 種，物販料金 (B) の種類は 58 種，チャージ (C) 料金の種類は 17 種である．なお，物販は簡単のため，料金の種類だけ商品の種類があると仮定する．

表 5.6 に用途別のエントロピー等の値を示す．X は特定の用途を示す．X =「交通」用途の場合，$H(U)=4.900$，$I(U;S)=3.085$ より，不明な値のある履歴からユーザが識別できる平均確率は $1/2^{H(U)}=0.033$ である．1 つの履歴レコードには，交通，物販，チャージのどれかひとつしか記録されていない．従って，1 つの履歴が判明した場合には，その確率は $1/2^{H(U)-I(U;S)}=1/2^{H(U;S)}=0.284$ まで上がる．$|U_X|$ はユニークユーザ数であり，交通 IC カードを用途 X で利用しているユーザの数である．例えば 31 人のユーザ中，交通 IC カードを「交通」用途で利用しているユーザは 31 人であるが,「物販」用途で利用しているのは 25 人である．また，$Pr(U,X)$ は用途 X の履歴を取得した場合に，データ中のあるユーザが特定される平均確率である．

$U'(S)=U'(B)=U'(C)$ の場合，I(U,X) の値で各用途ごとの情報量を比較することが可能であるが，本例は 3 用途とも $U'(X)$ が異なるため，$I(U,X)/H(U)$ の値で比較を行う．$I(U,X)/H(U)$ の値が最も大きいのは「物販」用途であり，1 つの履歴で $H(U)$ を 78.1%減らしている．このことから,「交通」「物販」「チャージ」の 3 用途の中では,「物販」用途の履歴の情報量が多い（危険である）といえる．

5.2.3 用途の相関

本節では，交通 IC カードデータのリスク分析をするため，用途間の関係について分析を行う．図 5.1 に交通利用回数とチャージ料金の関係を示し，図 5.2 に交通利用料金とチャージ料金の関係を示

表 5.6: 用途別のエントロピー等の値

	交通 (S)	物販 (B)	チャージ (C)	交通・物販 (S, B)
$H(U)$	4.900	4.338	4.736	4.412
$H(U\|X)$	1.814	0.948	3.256	0.182
$I(U;X)$	3.085	3.389	1.479	4.230
$Pr(U,X)$	0.284	0.518	0.105	0.881
$\|U_X\|$	31	25	29	31
ω_X	138	58	17	8004

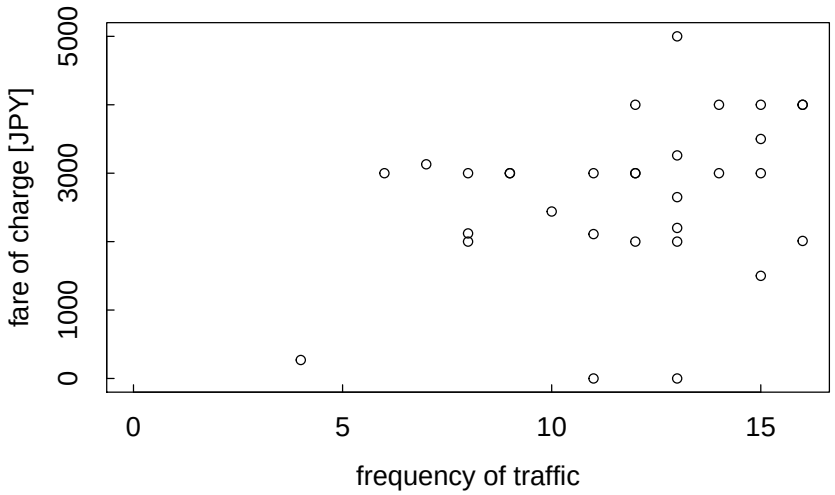

図 5.1: 交通利用回数とチャージ料金の散布図

す．相関係数は順に 0.469，0.315 であり，チャージ料金と交通利用回数・料金の間には弱い相関があることがわかる．よって，チャージ履歴の情報から交通履歴の情報を予測されるリスクがある．

交通履歴と物販履歴の関係を考える．表5.7 にユーザごとの物販料金の例 E_B を示す．E_S の u_1, u_2, u_3 と E_B の u_1, u_2, u_3 は同じユーザである．3.1 節と同様の手順で計算すると，$H(U) = 0.98$，$H(U|B) = 0.57$，$I(U;B) = 0.41$ が与えられる．

また，E_S と E_B から 1 つずつ履歴を取得することを仮定した集計表 $E_{S,B}$ を表 5.8 に示す．例として $E_{S,B}$ の場合，

$$Pr(u_1|s_1, b_1) = \frac{Pr(u_1|s_1)Pr(u_1|b_1)}{\sum_{i=1}^{n} Pr(u_i|s_1)Pr(u_i|b_1)} = \frac{4}{4+4} = \frac{1}{2}$$

と表すことができ，$H(U) = 1.19$，$H(U|S,B) = 0.46$，$I(U;S,B) = 0.73$ が与えられる．E_S,E_B,$E_{S,B}$ の各値を表 5.9 に示す．$I(U;x)$ の行より，

$$I(U;S) + I(U;B) = 0.89 > 0.73 = I(U;S,B)$$

である．このことから，交通と物販は独立ではないことがわかる．

交通 IC カードから取得した T の交通・物販用途を組み合わせた場合のエントロピー等の値を表

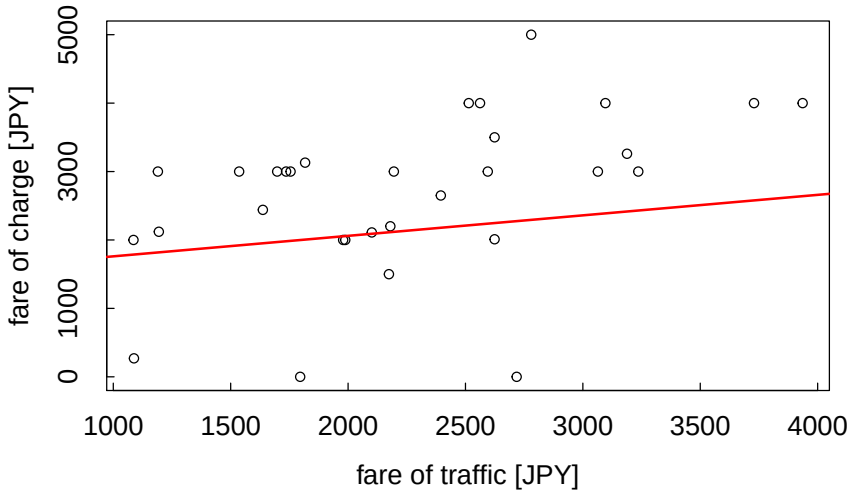

図 5.2: 交通利用料金とチャージ料金の散布図

表 5.7: ユーザごとの物販料金の例 E_B

ユーザ \ 物販料金	b_1	b_2	計	$Pr(U=u_i)$
u_1	2	0	2	2/7
u_2	1	3	4	4/7
u_3	0	1	1	1/7
$H(U\|B=b_i)$	0.92	0.31		
$Pr(B=b_i)$	3/7	4/7		

5.6 に示す．$I(U;S)$,$I(U,B)$ 等の結果と比較すると，

$$I(U;S)+I(U;B)=6.474>4.230=I(U;S,B)$$

となり，例と同様のことが言える．$\omega_X=8004$ は交通 (138 種類) と物販 (58 種類) の組み合わせの数である．交通と物販の履歴を取得した場合の相互情報量が大きくなっている．また，$Pr(U|X)$ は用途 X の 1 履歴が与えられたときのユーザの平均識別確率であるが，これも交通と物販の履歴を取得した場合が大きくなっている．よって，交通履歴と物販履歴を組み合わせることによる識別リスクが考えられる．

表 5.8: E_S と E_B から 1 履歴ずつ取得した場合の集計表 $E_{S,B}$

	s_1, b_1	s_1, b_2	s_2, b_1	s_2, b_2	s_3, b_1	s_3, b_2	計	$Pr(U = u_i)$
u_1	4	0	2	0	0	0	6	6/46
u_2	4	12	0	0	4	12	32	32/46
u_3	0	4	0	4	0	0	8	8/46
$H(U\|S = s_i, B = b_j)$	1	0.81	0	0	0	0		
$Pr(S = s_i, B = b_j)$	8/46	16/46	2/48	4/46	4/46	12/46		

表 5.9: E_S,E_B,$E_{S,B}$ の各値

$\backslash x$	s	b	s, b
$H(U)$	1.47	0.98	1.19
$H(U\|x)$	0.99	0.57	0.46
$I(U; x)$	0.48	0.41	0.73
$Pr(U\|x)$	0.50	0.67	0.73

第6章　まとめ

本稿では，現実的な攻撃者の想定とデータセットのリスク評価を行うため，攻撃者の持つ背景知識に注目した．データセットのある属性についての背景知識を持つ攻撃者の識別確率の期待値（平均識別確率）を用いることにより，リスク評価モデルを提案した．

3 章では購買履歴データ Online Retail Data Set に注目し，背景知識の異なる 10 タイプの攻撃者を想定し，これらの危険度を評価した．2 つの仮定（レコード数=ユーザ数，背景知識の独立性）をおくことにより，10 タイプの攻撃者の平均識別確率を近似する手法を提案したが，精度が非常に悪く，リスクを 2000 倍以上（真値：0.975，近似値：2413.9）に見積もってしまう場合もあった．

4 章では，平均識別確率を用いて 3 つの実データ（T_1:購買履歴データ，T_2:入院履歴データ，T_3:世帯収入データ）のリスク評価を行う実験を行った．その結果，T_1 では購買時属性が，T_2 では入院日数属性が，T_3 では年齢属性が最も危険である属性だと評価された．また，3 章で提案した近似手法を改善し，3 つの近似モデル（平均モデル，最小コストモデル，サンプリングモデル）を提案した．これらのモデルを用いて T_1, T_2, T_3 のリスクを評価した結果，データによっては最小コストモデル（計算コスト 0）でも，属性の危険度の順番を求めることができることが判明した．

5 章では，複数の属性から背景知識を得る攻撃者の危険度を近似する手法案として，エントロピーを用いたリスク評価手法を提案した．また，この手法を用いて交通 IC カードデータのリスク評価実験を行った結果，物販についての履歴の情報量が多く危険であることと，複数用途の情報を組み合わせることによってリスクがかなり大きくなることが判明した．

攻撃者モデルやリスク評価指標の想定は匿名加工を行うにあたっての大きな課題の一つであり，これを解決することによって匿名加工データのプライバシーリスク評価をこれまでより正しく評価できるようになるため，分野全体の研究が大きく前進する．本研究で提案した攻撃者想定やリスク評価モデルは，匿名加工技術の実用化に向けた大きな助けとなる．匿名化技術が実用化されれば，企業や組織はこれまで以上にビッグデータを安全に利活用できるようになり，ビッグデータ社会への大きな一歩となる．

参考文献

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

[2] 日本経済新聞, 「Suica 乗降履歴販売」失策の教訓 –パーソナルデータ活用 6 つの勘所–, https://www.nikkei.com/article/DGXNASFK1102K_R11C13A2000000/, refered in January 21, 2018.

[3] 個人情報保護委員会, 個人情報保護法について, https://www.ppc.go.jp/personalinfo/, refered in January 21, 2018.

[4] C. Dwork, “Differential privacy”, Proceedings of ICALP 2006, LNCS vol.4052, pp.1–12, 2006.

[5] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, “What is the Best Anonymization Method? - a Study from the Data Anonymization Competition Pwscup 2015”, Data Privacy Management Security Assurance (DPM2016), LNCS 9963, pp. 230 – 237, 2016.

[6] Klara Stokes, Vicen Torra, n-confusion: a generalization of k-anonymity, EDBT/ ICDT Workshops 2012: pp.211–215. (2012)

[7] Koot, M. R., Mandjes, M., van’t Noordende, G., and de Laat, C., “Efficient probabilistic estimation of quasi-identifier uniqueness”, In Proceedings of ICT OPEN 2011, 14–15, pp.119–126. (2011)

[8] Zhizhou Li, Ten H. Lai, “δ-privacy: Bounding Privacy Leaks in Privacy Preserving Data Mining”, DPM/CBT 2017, LNCS 10436, pp. 124-142, Springer. (2017)

[9] 早稲田 篤志, 野島 良, 盛合 志帆, 菊池 浩明,「良い仮名化　悪い仮名化」, 暗号とセキュリティシンポジウム 2017（SCIS-2017）, pp. 1–8, 2017.

[10] 南 和宏,「集計表秘匿における差分攻撃の考察」, 暗号とセキュリティシンポジウム 2017（SCIS-2017）, pp. 1–8, 2017.

[11] 南 和宏, 阿部穂日,「集計表セル秘匿問題の拡張によるデータ効用保持の有用性評価」, コンピュータセキュリティシンポジウム 2018（CSS-2018）, pp. 809 – 813, 2018.

[12] 濱田 浩気, 岡田 莉奈, 小栗 秀暢, 菊池 浩明, 中川 裕志, 野島 良, 波多野 卓磨, 正木 彰伍, 渡辺 知恵美,「匿名化アルゴリズムの公開・非公開による再識別容易性の比較」, 暗号とセキュリティシンポジウム 2018（SCIS-2018）, pp. 1–8, 2018.

[13] 山田 古都子，大圖 健史，石井 将大，田中 圭介,「大きい k に対する k-匿名化近似アルゴリズム」，暗号とセキュリティシンポジウム 2018（SCIS-2018），pp. 1–8，2018.

[14] 正木 彰伍,「攻撃者のモデル化を用いた軌跡情報の匿名性評価法」，コンピュータセキュリティシンポジウム 2017（CSS-2017），pp. 143 – 150，2017.

[15] 金沢 史明，岸野 徹,「特許出願からみた匿名化関連技術の技術動向 -平成 29 年度特許出願技術動向調査より-」，コンピュータセキュリティシンポジウム 2018（CSS-2018），pp. 906 – 912，2018.

[16] 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.

[17] 小栗 秀暢，黒政 敦史,「匿名加工情報の作成における攻撃者知識と安全性についての一考察」，コンピュータセキュリティシンポジウム 2017（CSS-2017），pp. 151 – 158，2017.

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

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

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

[21] 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.

[22] 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`, refered in December 17, 2018.

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

[24] 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`

[25] 伊藤 聡志,「乗降履歴データの有用性評価指標と匿名加工」，明治大学総合数理学部先端メディアサイエンス学科 2016 年度卒業論文，`https://windy.mind.meiji.ac.jp/paper/2016/bachelor/paper/ito.pdf`

謝辞

本稿は多くの方々のご指導・ご協力なくしては，完成しえないものである．

指導教官である明治大学総合数理学部の菊池浩明教授からは，著者が学部 1 年生の頃から現在までの 6 年間，多大なるご指導を賜った．両親は著者の学生生活を金銭的かつ精神的に，大いに支えてくれた．菊池研究室の同期や後輩たちは，著者の学生生活を楽しく，かけがえのないものにしてくれた．静岡大学の西垣正勝教授や大木哲史先生，東京電機大学の稲村勝樹先生をはじめとする他大学の先生方や，他研究室の学生の皆さんからは，合同発表会などを通じて，研究についての多くのご意見をいただいた．菊池研究室の社会人ドクターの新原功一氏，重本倫宏氏，仲小路博史氏，山口通智氏，ポスドクの黄緒平氏は，著者に研究者としてあるべき姿を示してくれた．学部の 1 期生である著者にとって，静岡大学の藤田真浩氏をはじめとする他大学の先輩方の存在は非常にありがたく，大いに面倒を見ていただいた．研究会やコンテスト等の際には，企業や研究所の方々から，研究や進路についての様々な助言を賜った．

著者の研究と学生生活を支えていただいた全ての方々に，心から感謝いたします．

研究業績

国際会議論文（査読あり）

1. Satoshi Ito, Hiroaki Kikuchi, “Risk of Re-Identification Based on Euclidean Distance in Anonymized Data PWSCUP2015”, Lecture Notes on Data Engineering and Communications Technologies, vol 7, proceedings of NBiS-2017, Springer, Cham, Canada, pp. 901–913, 2017.

2. Satoshi Ito, Reo Harada, Hiroaki Kikuchi, “Risk of Re-identification from Payment Card Histories in Multiple Domains”, 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA-2018), IEEE, Poland, pp. 934–941, 2018.

3. Satoshi Ito, Hiroaki Kikuchi, Hiroshi Nakagawa, “Attacker Models with a Variety of Background Knowledge of Payment History”, The 15th International Conference on Modeling Decisions for Artificial Intelligence (MDAI-2018), USB proceedings, Spain, pp. 178–189, 2018.

国内研究会

1. 伊藤 聡志，菊池 浩明，“ユークリッド距離を用いた再識別手法と PWSCup2015 の匿名加工データを用いた評価”，第 73 回コンピュータセキュリティ研究発表会（CSEC-73），pp. 1–8，2016.

2. 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”，暗号とセキュリティシンポジウム（SCIS-2017），pp. 1–8，2017.

3. 伊藤 聡志，原田 玲央，菊池 浩明，“乗降と物販履歴データの識別リスク分析と匿名加工の検討”，第 76 回コンピュータセキュリティ研究発表会（CSEC-76），pp. 1–8，2017.

4. 伊藤 聡志，菊池 浩明，中川 裕志，“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”，コンピュータセキュリティシンポジウム 2017（CSS-2017），pp. 1–8，2017