Computer Security Symposium 2019
21 - 24 October 2019

複数の価値算定法によるデータ評価値を高める高速項目削除方式

福岡 尊[1,a)]　山岡 裕司[1,b)]

概要：日本では，パーソナルデータのリスク算定手法として JO モデルが，価値算定手法として仮想市場モデルが，それぞれ考案されている．こういった価値やリスクの算定モデルから，表データの評価値を定めることができれば，例えばどのようにデータを部分的に削除すればよいかを決定するための指標になる．ただし，評価値を最大化するための部分的削除を行うには，一般には属性の組み合わせを全探索する必要があり，計算コストがかかる．本稿では，仮想市場モデルにより算定される価値と，JO モデルにより算定されるリスクの差を，表データの評価値として定める．本稿の貢献は，欠損が少ない表データに対して，この評価値を最大化する削除属性を決定することを目的とした，近似解を求める計算量の少ないヒューリスティックな手法の提案である．

キーワード：データ価値算定，リスク評価．

A fast method of deleting attributes of a dataset to improve a score derived from some methods of calculating the value of data

Takeru Fukuoka[1,a)]　Yuji Yamaoka[1,b)]

Abstract: In Japan, JO model and the hypothetical market model were introduced as methods of esimating the risk and the value of a personal dataset respectively. If we can define a score of a given dataset such calculation models, then that score, for example, may be used as an index concerning how to delete sub data of that dataset. In general, however, in order to decide which sub data should be deleted for maximizing that score, it is necessary to consider all sub data of that dataset, which will take a high calculation cost. In this proceeding, we define a score of a personal dataset as the difference between the value defined by the hypothetical market model and the risk defined by the JO model. The main contribution is a proposal of a heuristic fast method of deleting attributes of a given dataset to improve that score.

Keywords: Estimating the value of data, risk evaluation.

1. はじめに

戦略推進や収益増大化を行う上で，大量のパーソナルデータの利活用が注目されている．このパーソナルデータの取引の場として，データ取引市場や情報銀行といった，データ流通の仕組みが考案されてきている．例えば情報銀行 [3] は，個人が個人の権限の下で個人情報を情報銀行に託し，情報銀行はその個人情報を個人同意の下に流通させ，結果その報酬として個人にポイント贈呈などのサービスを与えるといった構造をなす．この情報銀行にはすでに認定制度ができており，流通させたデータを企業が漏洩した場合，その責任を情報銀行も負うというモデルになっている．そのため情報銀行は，保持するパーソナルデータの価値とリスクを，どちらも管理する必要がある．

データをスコアリングして管理することを考えたとき，仮にパーソナルデータの価値とリスクが計算できれば，そ

[1] 富士通研究所
FUJITSU LABORATORIES LTD.
[a)] fukuoka.takeru@fujitsu.com
[b)] yamaoka.yuji@fujitsu.com

の差額である「価値」–「リスク」を，データの評価値として算出することが可能となる．この評価値が大きければ大きいほど，リスクに比べ価値が高いデータであることがわかる．そのためこの評価値は，パーソナルデータの管理手法を判断する一つの指標として機能することが期待される．

本稿で扱うパーソナルデータは，行により個人を，列により属性を表す，個人の属性値をセルの成分とするような，欠損を許す表データである（表 1 を参照）．この場合，評価値である「価値」–「リスク」は，属性を削除することで増加しうる．本稿で取り扱う問題は，「どの属性らを削除すれば，評価値が最大化されるか」という問いである．この問題を一般に解くには，削除する属性の組み合わせを全探索しなければならないため，時間がかかる．本稿の主な目標は，この問題を効率よく求める手法の探索である．

本稿では，パーソナルデータの価値を算定するモデルとして一般財団法人日本情報経済社会推進協会（以下，JIPDEC）の作成した仮想市場モデルを，リスクを算定するモデルとして NPO 日本ネットワークセキュリティ協会（以下，JNSA）の作成した JO モデルを，それぞれ用いる．JO モデルは，リスク算定モデルとして広く使われているモデルであり，例えば一般社団法人日本サイバーセキュリティ・イノベーション委員会（以下，JCIC）のサイバーリスク指標モデルなどでも用いられている [2]．一方，JIPDEC の作成した仮想市場モデルは，パーソナルデータの仮想的な市場価値を算定するモデルであり，専門家に対するヒアリングを行うことでその妥当性を検証している [1]．またこれらは，個人情報の価値とリスクを円単位で算出するため，互いに比較可能である．そこで「仮想市場モデルでの値」–「JO モデルでの値」として評価値を導入する．本稿の貢献は，この導入された評価値に対して，どの属性を削除すれば評価値が最大になるかという問題を解くための，近似解を与える高速なアルゴリズムの提案である．

2. JO モデルと仮想市場モデル

本章ではまず，JO モデルと仮想市場モデルを復習する．JO モデルは，個人情報のリスク算定モデルとして，2003 年に JNSA によるワーキンググループにおいて，IT，保険，セキュリティなどの専門家により作られた [4]．この JO モデルは，パーソナルデータが漏洩した際，訴訟された場合の予想賠償金額を，円単位で算出するモデルである．この JO モデルの算出する方式を参考に，パーソナルデータの仮想市場価値を算定したモデルが，仮想市場モデルである．本章では，これらの定義を振り返ったのち，仮想市場モデルと JO モデルでの値の差として，パーソナルデータの評価値を定める．

2.1 パーソナルデータ

JO モデルや仮想市場モデルの定義を行う前に，対象とするパーソナルデータを定める．本稿において，**パーソナルデータ**は，個人，および（氏名，年齢，位置情報といった）属性により行と列が添え字づけられた，成分を文章や数字とする行列である．以下の表 1 に一例を示す：

表 1 パーソナルデータの例

	位置情報	購買記録	生体記録	診療記録
a	新宿 13:13	湿布 300 円	体温 36.4	腰痛
b	渋谷 20:54	清涼飲料水 100 円	—	—
c	池袋 09:30	—	体温 38.2	風邪
d	上野 16:00	—	—	頭痛

2.2 JO モデルの定義

JO モデルは，一人分のパーソナルデータの漏洩時予想損害賠償額を，以下の 5 つの値の積により定めている：

「機微情報度」
×「基礎情報価値 ($= 500$ 円)」×「本人特定容易度 ($\in \{1,3,6\}$)」
×「社会的責任度 ($\in \{1,2\}$)」×「事後対応評価 ($\in \{1,2\}$)」.

よって表形式のパーソナルデータの漏洩時予想損害賠償額は，各個人に対する JO モデルでの値を計算し，その和をとることで算定される．以下，[4] における，この各 5 つの値の決め方を確認する．

2.2.1 機微情報度

機微情報度とは，漏洩した個人情報に含まれる機微情報の量を表している．この量を定義するためには，各属性に対して，その属性が漏洩したときの精神的苦痛レベルと，経済的苦痛レベルを定める必要がある．厳密にいえば，属性数が m であるとき，**精神的苦痛レベル**および**経済的苦痛レベル**は，各属性（を表す列番号）に対して正の整数を返す写像 $p, e\colon \{1,\dots,m\} \to \mathbb{Z}_{>0}$ のことである．これら p, e のとる値を決める数理的プロセスはないものの，代表的な属性に対しては，JO モデルによってすでに決定されている [4].

この精神的苦痛レベルと経済的苦痛レベルを用いることで，情報機微度は以下のように定められる．一人分のパーソナルデータに対して，そのデータの持つ属性を $1,\dots,m$ により添え字付ける．欠損していない属性箇所を $J \subset \{1,\dots,m\}$ としたとき，**情報機微度** S は以下で定義される：

$$S := 10^{\max_{j\in J} p(j)-1} + 5^{\max_{j\in J} e(j)-1}. \qquad (1)$$

ここで，$p(j)$ は j の精神的苦痛レベル，$e(j)$ は j の経済的苦痛レベルを表す．

計算例 2.1. 表 1 の属性たちを例に考える．左から順に番号づけると，$p(1), p(2), p(3), p(4)$ (resp. $e(1), e(2), e(3), e(4)$) はそれぞれ「位置/時刻情報」，「購買記録」，「生体記録」，

「診療記録」の精神的苦痛レベル (resp. 経済的苦痛レベル) に対応し，これらは JO モデルによってそれぞれ $1,1,2,3$ (resp. $1,2,1,1$) と定められている [4].

そこで，表 1 における $S(3)$ を計算してみる．3 行目のデータは購買記録のみ欠損しているため，欠損していない箇所は $\{1,3,4\}$ である．すると，$(p(1),p(3),p(4)) = (1,2,3)$，$(e(1),e(3),e(4)) = (1,1,1)$ であるため，$\max\{p(1),p(3),p(4)\} = 3$，$\max\{e(1),e(3),e(4)\} = 1$ であるから，$S(3) = 10^{3-1} + 5^{1-1} = 101$ と求められる．

2.2.2 その他の要素

基礎情報価値は，一律 500 円として定められている．

本人特定容易度は，漏洩した個人情報から個人をどれだけ特定しやすいかを表す数値であり，その個人のもつ欠損していない属性が「氏名」「住所」を含んでいれば 6，「氏名」または「住所+電話番号」のどちらか一方のみを含んでいる場合は 3，それ以外は 1 として定義する．

社会的責任度は，そのデータを取り扱っている企業が，医療，金融・信用，情報通信等の業種ならば 2，知名度の高い大企業や公的機関ならば 2，それ以外の場合は 1 の値をとる．

事後対応評価は，実際に漏洩した際，不適切な対応をとったなら 2，そうでないならば 1 の値をとる．実際に漏洩する前は，この値は 1 である．

計算例 2.2. これらの定義をもとに，表 1 の JO モデルでの値を求める．ただし，事後対応評価，社会的責任度は，それぞれ 1 とする．さらに，属性には氏名も住所も含まれないので，本人特定容易度も 1 である．個人を上から順に 1,2,3,4 と番号付けし，$i \in \{1,2,3,4\}$ に対する欠損していない箇所を J_i とすれば，$J_1 = \{1,2,3,4\}$, $J_2 = \{1,2\}$, $J_3 = \{1,3,4\}$, $J_4 = \{1,4\}$ である．例 2.1 で見たように精神的苦痛レベルと経済的苦痛レベルが定まっているので，番号 $1,2,3,4$ である個人の情報機微度 (式 (1)) は，$105, 6, 101, 101$ と定まる．よって JO モデルでの値はそれぞれ 52500 円，3000 円，50500 円，50500 円となり，表全体の JO モデルでの値は 156500 円である．

2.3 仮想市場モデル

この JO モデルの手法をもとに，JIPDEC はパーソナルデータの価値を円単位で推定する，JO モデルとは異なるモデルを構成した [1]．本稿では，このモデルを**仮想市場モデル**と呼ぶ．対象とするデータは，位置情報を含むパーソナルデータであり，位置情報にどのようなデータが付帯されたときに，市場価値がいくらになるかを仮想的に算定することを目的としている．仮想市場モデルは，一人分の位置情報を付帯するパーソナルデータの情報価値を，以下の 3 つの値の積として定めている：

「情報価値度」

×「(位置情報の) 基礎情報価値 (= 1 円)」×「匿名化度」.

よって表形式のパーソナルデータの仮想的な市場価値は，各個人に対する仮想市場モデルでの値を計算し，その和をとることで算定される．

以下，[1] における，一人分のパーソナルデータに対する，この各 3 つの値の決め方を確認する．

2.3.1 情報価値度

仮想市場モデルにおける情報価値度は，JO モデルにおける機微情報度とほとんど同様に定義される．まず，一人分の位置情報を含むパーソナルデータの持つ各属性に対し，位置情報以外の属性に対して，**受容性 (receptivity)** と**有用性 (usefulness)** を定める．これらは，位置情報以外の属性を $1,\dots,m$ と添え字付けたとき，写像 $r, u\colon \{1,\dots,m\} \to \mathbb{Z}_{>0}$ として定義される．ただし，代表的な属性に対しては，仮想市場モデルによって既にその受容性と可用性が定まっており，それは $1,2,3$ のいずれかの値をとる [1]．さらに $J \subset \{1,\dots,m\}$ を欠損していない属性箇所としたとき，**情報価値度** V は以下で定義される：

$$V := 8^{\max_{j\in J} r(j)-1} + 6^{\max_{j\in J} u(j)-1}. \qquad (2)$$

ここで $r(j)$ は j 番目の属性の受容性，$u(j)$ は有用性を表す．

注意 2.3. 個人情報として，識別子と位置情報しかないデータも当然考えられる．このデータの価値は，そのまま基礎情報価値であるべきと考える．そこで本稿では，識別子と位置情報しかないデータの情報価値度は，特別に 1 として定める．

2.3.2 その他の要素

基礎情報価値は，JO モデルとは異なり，1 円と定められている．この値段は，ブラックマーケットにおける取引金額，個人情報漏洩事件における名簿業者への売却価格などを参考に設定された．

匿名化度については，詳細な定義は省かれているが，原典である [1] による研究報告書には，「経済産業省の情報大航海プロジェクト（平成 19 年度～21 年度）では，匿名化前後のレコード間でどのぐらい情報が異なっているかを評価（情報損失）することで，匿名化情報の有用性を示した．データ流通する際は，識別情報を匿名化（切り落としまたは仮名化）を施す．この状態を基準値の『1』とした.」と記されている．

計算例 2.4. これらの定義をもとに，表 1 の仮想市場モデルでの値を計算する．まず，表 1 は位置情報が付帯されているデータであるため，仮想市場モデルによって計算できることに注意する．位置情報でない属性である「購買記録」,「生体記録」,「診療記録」を左から順に $1,2,3$ と添え字づけると，それらの受容性 $r(1), r(2), r(3)$ (resp. 有用性 $u(1), u(2), u(3)$) はそれぞれ $2,3,3$ (resp. $3,3,3$) となる．

個人を上から順に 1,2,3,4 と番号付けし，$i \in \{1,2,3,4\}$ に対する欠損していない箇所を J_i とすれば，$J_1 = \{1,2,3\}$, $J_2 = \{1\}$, $J_3 = \{2,3\}$, $J_4 = \{3\}$ であるので，番号 $1,2,3,4$ である個人の情報価値度はそれぞれ $100, 44, 100, 100$ と定まる．匿名化度は 1，基礎情報価値は 1 円であるため，仮想市場モデルでの値は 344 円となる．

2.4 JO モデルと仮想市場モデルの比較

以上のように，JO モデルと仮想市場モデルは，どちらもパーソナルデータの価値算定モデルである．しかしながら，その価値評価のアプローチは大きく異なる．実際，JO モデルは，パーソナルデータが漏洩した際の想定損害賠償額というリスク評価であり，仮想市場モデルは，パーソナルデータが仮に取引された場合の価格という市場価値評価である．また，値段のスケールも大きく異なることも特徴である．実際表 1 は，JO モデルでの金額は 156500 円，仮想市場モデルでの金額は 344 円であった．ただし，JO モデルでの金額は，訴えられた場合の値段であることから，実際には訴訟参加率も考慮に入れた形で被害額が推定されるケースもある [5], [6]．特に JNSA の 2010 年のセキュリティ被害調査 WG の活動報告 [5] では，訴訟参加率を 0.1% として被害額計算を行っている．この設定を踏襲した場合，表 1 の JO モデルでの値段に訴訟参加率を加味した値は 156.5 円となり，両者の値段のスケールはほぼ等しくなる．

3. 仮想評価値最大化問題

3.1 仮想評価値：仮想市場モデルと JO モデルの差

本稿では，パーソナルデータの**仮想評価値**を

$$\text{「仮想市場モデルでの値」} - \text{「JO モデルでの値」} \times 0.1\%. \tag{3}$$

と定義する．ここで，0.1% は訴訟参加率である．この値が正 (resp. 負) であれば，仮想市場での価値が想定リスク評価値を上回って (resp. 下回って) いる．この評価値は，例えばデータのどの部分を破棄すべきか，それとも保持すべきかの指標を与えるなどの点で，データ運用の一つの指標として期待できる．

計算例 3.1. 表 1 の JO モデルでの値は 156500 円，仮想市場モデルでの値は 344 円であった．よって評価値は，$344 - 156.5 = 187.5$ 円となる．以下，データを部分的に削除したときの，評価値の移り変わりを見ていく．

- 表 1 から属性「診療記録」を削除したパーソナルデータは，仮想市場モデルでの値が 245 円，JO モデルでの値が 16500 円となるので，評価値は 228.5 円となり，削除する前よりも評価値が増加する．
- 表 1 から属性「生体記録」と「購買記録」を削除したパーソナルデータは，仮想市場モデルでの値が 301 円，JO モデルでの値が 152500 円となるので，評価値は 148.5 円となり，削除する前よりも評価値が低下する．

3.2 仮想評価値最大化問題

本稿において問いたい問題は，以下である．

問題 3.2. 与えられたプライバシーデータに対して，どの属性を削除すれば，仮想評価値を最大化できるか？

記号 3.3. 以下では，個人数 n，属性数 m のプライバシーデータを考える．対応する行列は n 行 m 列である．そこで $\{1,\dots,m\}$ を属性の集まり，$\{1,\dots,n\}$ を個人の集まりとしてみなす．$\{1,\dots,m\}$ の部分集合 J に対し，J に対応する列のみを残した行列の，仮想評価値を $v(J)$ とおく．また，argmax で最大値集合を表す．

3.3 既存解法その 1：総当たり法

問題 3.2 は離散最適化問題である．そのため，属性のなす集合 $\{1,\dots,m\}$ の部分集合を全て考えるという**総当たり法**によって，最適解を求めることができる．より正確には，以下の解法で問題 3.2 の最適解を求めることができる．

解法 3.4 (総当たり法)**.** 記号 3.3 を用いる．まず $\{1,\dots,m\}$ の任意の部分集合 J に対し，J に対応する列のみを残した行列の仮想評価値 $v(J)$ を計算する．そこで最大値集合 $\mathrm{argmax}_{J\subset\{1,\dots,m\}} v(J)$ の要素を一つ $J_{\max}$ としてとると，$J_{\max}$ は仮想評価値を最大化する属性の集まりとなる．この解法の計算量は $O(n2^m)$ である． ∎

しかしながらこの解法は，属性数が増えた際に計算量が爆発的に増加するという問題を抱えている．

3.4 既存解法その 2：局所探索法

また，問題 3.2 の近似解を求めるために，既存のメタヒューリスティックな解法を用いることも考えられる．代表的な解法の一つとして，本稿では局所解探索の一例を取り上げる．

局所解探索を用いるには，近傍を定義しなければならない．本稿では，以下のような Hamming 距離の類似を用いる．すなわち，$\{1,\dots,m\}$ の部分集合 $J_1, J_2 \subset \{1,\dots,m\}$ に対して，$h(J_1,J_2) = \#(J_1 \setminus J_2) + \#(J_2 \setminus J_1)$ とおく．そして $J \subset \{1,\dots,m\}$ の近傍 $N(J)$ を

$$N(J) := \{J' \subset \{1,\dots,m\} \mid h(J,J') = 1\}$$

として定義する．この設定の下，以下のように探索法を定義する．

解法 3.5 (局所探索法)**.** 記号は記号 3.3 のものを用いる．

(1) 各属性 $j \in \{1,\dots,m\}$ に対して，その属性の仮想評価値が小さい順に $\{1,\dots,m\}$ を並び替える．現在状態 $S \subset \{1,\dots,m\}$ をランダムに用意する．さらにタブー

リストを表す集合を T とし，初期状態として $T = \{S\}$ とする．

(2) S との Hamming 距離が 1 である集合 $N(S)$ を用意する．$N(S)$ の各要素 S' に対して $v(S')$ を計算し，その最大値を V' とする．このとき $\mathrm{argmax}_{S' \in N(S)} v(S') = \{S' \in N(S) \mid v(S') = V\}$ である．

(3) $V' \geq v(S)$ かつ $\mathrm{argmax}_{S' \in N(S)} v(S') \not\subset T$ であるときは，集合 $\mathrm{argmax}_{S' \in N(S)} v(S') \setminus T$ の各要素 S' について $t(S')$ を以下で定める：

- $S \subset S'$ であるときは，$S' - S = \{s'\}$ について $t(S') = s'$ とする．
- $S' \subset S$ であるときは，$S - S' = \{s'\}$ について $t(S') = m - s'$ とする．

$t(S')$ が大きいということは，属性の仮想評価値が大きいものを付け加えている，もしくは属性の仮想評価値が小さいものを取り除いているということを意味する．そこで，$t(S')$ ができるだけ大きい値をとる S' をとり，T に S' を加え，S を S' に置き換え，(2) へ戻る．

(4) $V' < v(S)$，もしくは $\mathrm{argmax}_{S' \in N(S)} v(S') \subset T$ であるときは，S を出力し終了する．

タブーリストとして T を導入し，今までに探索した部分集合を記録することで，無限ループに陥らないようにしている．この手法によって状態を探索すると，少なくとも局所解にたどり着くことはできる．またアルゴリズムの各ステップは計算量は $O(nm + m \log m)$ で抑えられる．ステップ数を評価することは難しいため，実装の際には，タブーリスト T の長さに閾値を設ける可能性もありうる．

しかしながら，与えられているプライバシーデータによっては局所解が多く存在するため，その場合は最適解から遠い結果を与えやすい解法であることが予想される．

4. 提案手法

本稿の主な貢献は，JO モデルと仮想市場モデルの特性を活かすことで，より仮想評価値問題に特化したヒューリスティック解法を提案することにある．まず § 4.1 で今回の問題を列削除問題として厳密に定式化する．その後，§ 4.2 で，補助問題としてセル削除問題を導入し，その解法を与える．その解法を用いて，§ 4.3 で，提案方式を定式化する．

4.1 列削除問題の定式化

まず，今回取り扱っている問題を少し簡略化したものを，正確に定式化する．

問題 4.1 (列削除問題)．n, m, l を正の整数とする．写像 $X\colon \{1, \ldots, n\} \to \mathfrak{P}(\{1, 2, \ldots, m\})$ を固定する ($\mathfrak{P}$ でべき集合を表す)．さらに写像 $A\colon \{1, \ldots, m\} \to \{1, 2, \ldots, l\}^4$ を考える．各集合 $K \subset \{1, 2, \ldots, l\}^4$ に対して，

$$\mu(K) := \left(\max_{k \in K} \mathrm{pr}_1(k), \max_{k \in K} \mathrm{pr}_2(k), \max_{k \in K} \mathrm{pr}_3(k), \max_{k \in K} \mathrm{pr}_4(k) \right)$$

と定義する．ここで $\mathrm{pr}_i\colon \{1, 2, \ldots, l\}^4 \to \{1, 2, \ldots, l\}$ は第 i 成分への射影である．さらに t_1, t_2, t_3, t_4 を定数，さらに c を定数とし，

$$v(k_1, k_2, k_3, k_4) := (t_3^{k_3 - 1} + t_4^{k_4 - 1}) - c(t_1^{k_1 - 1} + t_2^{k_2 - 1}) \tag{4}$$

と定義する．このとき，$\sum_{i=1}^{n} v(\mu(A(J \cap X(i))))$ を最大化する $J \subset \{1, \ldots, m\}$ を見つけよ．

注意 4.2. 上記の問題がなぜ今回の問題の定式化となるのかを説明する．まず，n 行 $(m+1)$ 列のパーソナルデータを考え，個人を $1, \ldots, n$，属性を $1, \ldots, m, m+1$ で添え字付ける．ただし $(m+1)$ 番目は常に位置情報であるとし，その情報には欠損値はないと仮定する．さらに簡略化のため，全ての個人の本人特定容易度は 1 であるとする．例えば，「氏名」と「住所」がないデータだったとする．

上記の設定の下，各属性に対して精神的苦痛レベル，経済的苦痛レベル，可用性，有用性を与える写像をそれぞれ $p, e, r, u\colon \{1, \ldots, m\} \to \{1, 2, \ldots, l\}$ とする．JO モデルも仮想市場モデルも，現在は $l = 3$ までの値しかとらないが，将来的により細かいレベル分けが行われることも考え，一般化する．これにより自然に写像 $A\colon \{1, \ldots, m\} \to \{1, \ldots, l\}^4$ が誘導される．

v の定義式において，$(t_1^{k_1 - 1} + t_2^{k_2 - 1})$ は仮想市場モデルにおける情報価値度，$(t_3^{k_3 - 1} + t_4^{k_4 - 1})$ は JO モデルにおける機微情報度を表している．従来では $t_1 = 8$, $t_2 = 6$, $t_3 = 10$, $t_4 = 5$ であったが，将来的な変更も考え，一般化する．仮想市場モデルや JO モデルは，情報価値度および機微情報度を定数倍することで算出していたため，定数 c を状況に応じて設定することで仮想評価値（をスケーリングしたもの）を表現できる．例えば $t_1 = 8$, $t_2 = 6$, $t_3 = 10$, $t_4 = 5$, $c = 1/2$ とすることで，$\sum_{i=1}^{n} v(\mu(A(X(i))))$ は既存の設定における仮想評価値を表す．$J \subset \{1, \ldots, m\}$ に対する $\sum_{i=1}^{n} v(\mu(A(X(i) \cap J)))$ は，J 以外の属性をパーソナルデータから削除した際の仮想評価値を表している．

4.2 セル削除問題

上述の問題 4.1 はもちろん，解法 3.4 により解くことができ，計算量は $O(n2^m)$ であった．一方で，この問題を n, m に関する多項式時間で解くことはかなり難しいと予想している．そこで以下のより簡単な問題を考える：

問題 4.3 (セル削除問題)．問題 4.1 と同じ設定の元で，$\sum_{i=1}^{n} v(\mu(A(J_i \cap X(i))))$ を最大化する $J_i \subset \{1, \ldots, m\}$ を各 i に対して見つけよ．

この問題は，各行に対して，仮想評価値を最大化するた

めの削除すべき属性を選択すればよいだけなので，以下のように解くことができる．

記号 4.4. $k, k' \in \{1,\dots,l\}^4$ について，$k \le k' :\iff \forall i \in \{1,2,3,4\}, \mathrm{pr}_i(k) \le \mathrm{pr}_i(k')$ と定義する．また，有限集合 S に対して，その濃度を $\#S$ で表す．

解法 4.5 (問題 4.3 の解法)**.** 前処理として，$j \in \{1,\dots,m\}$ に対する $v(j)$ の値によって属性 $\{1,\dots,m\}$ をソートしなおしておく．ただし，$j < j'$ ならば $v(j) \le v(j')$ とする．各行 $i \in \{1,\dots,n\}$ に対して，J_i と J_i' を以下のアルゴリズムにより定義する．

(1) まず $A\colon \{1,\dots,m\} \to \{1,\dots,l\}^4$ により，$K := A(X(i))$ とする．

(2) $\mu = \mu(K)$ とし，集合 $S = \{k \in \{1,\dots,l\}^4 \mid k \le \mu\}$ を考える．

(3) 集合 S の要素 $k = (k_1,k_2,k_3,k_4)$ を $v(k) = t_3^{k_3-1} + t_4^{k_4-1} - c(t_1^{k_1-1} + t_2^{k_2-1})$ の大小によってソートし，$(k^1,\dots,k^{\#S})$ とする（ただし $v(k^i) \ge v(k^{i+1})$）．

(4) 以下のアルゴリズムを走らせることで，部分集合 $M \subset K$ を得る．

(a) 初期値 $x = 1$ とする．
(b) $K_x' := \{k \in K \mid k \le k^x\}$ とし，$\mu_x' := \mu(K_x')$ とする．
(c) 仮に $\mu_x' = k^x$ ならば，$M = K_x'$ とする．そうでないならば，x を $x+1$ に取り換え，(b) に戻る．

(5) $J_i = A^{-1}(M)$ と置く．さらに $J_i = \{j_i^1,\dots,j_i^l\}$ とおき，以下のアルゴリズムによって $J_i' \subset J_i$ を定める．

(a) 初期値として $J_i' = J_i$ とおく．
(b) J_i' の元を小さい順に並べたものを $(j_i^1,\dots,j_i^l)$ とし，各 $k \in \{1,\dots,l\}$ に対し $(J_i')_k := J_i' \setminus \{j_i^k\}$ とする．
(c) ある k で $v(\mu(A((J_i')_k \cap X(i)))) = v(\mu(A(J_i \cap X(i))))$ を満たすものがあれば，そのような k の内最も小さいものをとり，J_i' を $(J_i')_k$ に取り換え，(b) に戻る．
(d) 任意の k で $v(\mu(A((J_i')_k \cap X(i)))) < v(\mu(A(J_i \cap X(i))))$ が成り立つならば，J_i' を出力する．

(6) J_i と J_i' はどちらも問題 4.3 の解となるので，それらを出力する．

注意 4.6. 本方式の計算量は，高々 $O(nl^4 \log l)$ で抑えられる．また，各行 i に対する問題 4.3 の最適解 J_i と J_i' の意味は，J_i が「削除するセルを極力減らした解」であるのに対し，J_i' は「削除するセルを極力増やした解の一つ」である．二つ用意した目的は，後述する本稿で提案する解法の正答率を上げることにある．

4.3 列削除問題のヒューリスティックな解法

本稿では，このセル削除問題の解法を活かして，欠損値が比較的少ないパーソナルデータに対してうまく働くような，ヒューリスティックな解法を提案する．ポイントは，セル削除による解によって「どの列を残すべきか」を表す指標を導入できることにある．

解法 4.7 (列削除問題のヒューリスティックな解法)**.** 列数が少ない場合は全探索で現実的な時間内に解けるので，あらかじめ全探索解法を許容する列数の閾値 s を固定する．この s を固定したうえで，以下のアルゴリズムを走らせる．

(1) セル削除方式 (解法 4.5) によってセル削除問題をとき，各 $i \in \{1,\dots,n\}$ に対し残すべき箇所 J_i と J_i' が求まる．

(2) 各 $j \in \{1,\dots,m\}$ に対して，

$$f(j) = \frac{\#\{i \in \{1,\dots,n\} \mid j \in J_i\}}{\#\{i \in \{1,\dots,n\} \mid j \in X(i)\}}$$

$$f'(j) = \frac{\#\{i \in \{1,\dots,n\} \mid j \in J_i'\}}{\#\{i \in \{1,\dots,n\} \mid j \in X(i)\}}$$

とおく．これは j 列目の非欠損値の数を分母，解法 4.5 を施した後の非欠損値の数を分子とする分数である．この $f(j), f'(j) \in [0,1]$ は，第 j 列を削除するべきかの度合を示す指標であり，小さければ小さいほど削除するべき行であることがわかる．

(3) $(|f(j) - 1/2|, -f(j))$ が辞書式順序で大きい順に並べたものを $(j_1,\dots,j_m)$ とし，$f(j) - 1/2$ が負なら削除，正なら残すという操作を j_1 から j_{m-s} まで繰り返す．辞書式順序を考える理由は，解 $(J_i)_{1\le i\le n}$ はできるだけセルを削除しない解であるから，削除されている，すなわち $f(j)$ が小さい方を優先すべきだからである．同様に，$(|f'(j) - 1/2|, f'(j))$ が辞書式順序で大きい順に並べたものを $(j_1',\dots,j_m')$ とし，$f(j) - 1/2$ が負なら削除，正なら残すという操作を j_1' から j_{m-s}' まで繰り返す．しかしここで，$(J_i')_{1\le i\le n}$ はできるだけセルを削除する解であるから，削除されていない，すなわち $f(j)$ が大きい方を優先すべきなので，その部分のみ変更してある．

(4) 残りの列 $\{j_{m-s+1},\dots,j_m\}$ および $\{j'_{m-s+1},\dots,j'_s\}$ について全探索にて列削除問題を解き，残すべき列番号のなす集合 $J, J' \subset \{1,\dots,m\}$ を得る．そして $v(J)$ と $v(J')$ の値が大きい方をアウトプットとして出力する．

計算量は $O(l^4 \log l + mn + (m-s)\log(m-s) + n \cdot 2^s)$ である．

4.4 本解法 4.7 の実施例

$s = 3$ 列までは全探索可能であると仮定する．パーソナ

ルデータの模式図として，以下の表 (5) を考える．

$$\begin{pmatrix} & \text{列 1} & \text{列 2} & \text{列 3} & \text{列 4} & \text{列 5} & \text{列 6} \\ \text{行 1} & \circ & \circ & \circ & \circ & \circ & \circ \\ \text{行 2} & \circ & \circ & - & \circ & \circ & \circ \\ \text{行 3} & \circ & \circ & \circ & - & \circ & \circ \\ \text{行 4} & \circ & \circ & \circ & \circ & \circ & - \\ \text{行 5} & \circ & \circ & - & - & \circ & \circ \end{pmatrix} \tag{5}$$

ここで，ハイフンは欠損値，丸は欠損値でないことを表す．v の定義式 (式 (4)) に現れる定数はそれぞれ，$t_1 = 5$, $t_2 = 2$, $t_3 = 3$, $t_4 = 4$, $c = 1$ とする．さらに $A(1), A(2), A(3), A(4), A(5), A(6)$ を $(2,2,1,1)$, $(2,2,1,2)$, $(2,2,3,2)$, $(2,3,1,3)$, $(3,3,1,3)$, $(3,3,3,2)$ とおく．

(1) セル削除問題の解法 4.7 を表 (5) に適応すれば，以下の二つの答えを得る：

$$\begin{pmatrix} J_1 \\ J_2 \\ J_3 \\ J_4 \\ J_5 \end{pmatrix} = \begin{pmatrix} \circ & \circ & \circ & \circ & - & - \\ \circ & \circ & - & \circ & - & - \\ \circ & \circ & \circ & - & - & - \\ \circ & \circ & \circ & \circ & - & - \\ - & - & - & - & - & - \end{pmatrix}$$

$$\begin{pmatrix} J'_1 \\ J'_2 \\ J'_3 \\ J'_4 \\ J'_5 \end{pmatrix} = \begin{pmatrix} - & - & \circ & \circ & - & - \\ - & - & - & \circ & - & - \\ - & - & \circ & - & - & - \\ - & - & \circ & \circ & - & - \\ - & - & - & - & - & - \end{pmatrix}$$

(2) どの列を削除するべきかの指標 $f(j)$ および $f'(j)$ はそれぞれ $(4/5, 4/5, 3/3, 3/3, 0/5, 0/4)$, $(0/5, 0/5, 3/3, 3/3, 0/5, 0/4)$ となる．

(3) よって，$(|f(j) - 1/2|, -f(j))$ が辞書式順序で大きい順に並べると，$j_1 = 5$, $j_2 = 6$, $j_3 = 3$ となる．したがって 5, 6 列目は削除し，3 列目は残す．

$$\begin{pmatrix} \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & - & \circ & \circ & \circ \\ \circ & \circ & \circ & - & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & - \\ \circ & \circ & - & - & \circ & \circ \end{pmatrix} \to \begin{pmatrix} \circ & \circ & \circ & \circ & - & - \\ \circ & \circ & - & \circ & - & - \\ \circ & \circ & \circ & - & - & - \\ \circ & \circ & \circ & \circ & - & - \\ \circ & \circ & - & - & - & - \end{pmatrix}.$$

残りの列 $\{1, 2, 4\}$ に関して全探索をかけると，答えとして $J = \{3, 4\}$ を得る．

同様に，$(|f'(j) - 1/2|, f'(j))$ が辞書式順序で大きい順に並べると，$j_1 = 3$, $j_2 = 4$, $j_3 = 1$ となる．したがって 3, 4 列目は残し，1 列目は削除する．

$$\begin{pmatrix} \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & - & \circ & \circ & \circ \\ \circ & \circ & \circ & - & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & - \\ \circ & \circ & - & - & \circ & \circ \end{pmatrix} \to \begin{pmatrix} - & \circ & \circ & \circ & \circ & \circ \\ - & \circ & - & \circ & \circ & \circ \\ - & \circ & \circ & - & \circ & \circ \\ - & \circ & \circ & \circ & \circ & - \\ - & \circ & - & - & \circ & \circ \end{pmatrix}.$$

残りの列 $\{2, 5, 6\}$ について全探索をかければ，答えとして $J' = \{3, 4\}$ を得る．

(4) 今回は J，J' が一致したので，$\{3, 4\}$ が出力となる．

4.5 本解法 4.7 の考察

本解法 4.7 の特徴は，セル削除解法 (= 解法 4.5) を援用することで，j 番目の列を残すべきかを表す指標 $f(j), f'(j)$ を解法 4.7 (2) にあるように導入する点にある．セル削除解法は，欠損のないデータに対しては列削除問題 (= 問題 4.1) に対する解を与えるため，欠損値が比較的少ない表データに対しては，解法 4.7 は有効に働くと期待できる．

5. 実験

従来の局所探索法 (=解法 3.5) と，今回提案した探索法 (=解法 4.7) とを比較するために，実験を行う．同程度の計算をさせたとき，どちらがよりよい解を出すことができるかをもってして，二つの解法の優劣を比較する．特に，行列の欠損値の割合が少ない場合を中心に観察する．

5.1 設定

n 人の個人からなる m 個の属性を持つパーソナルデータについて，JO モデルおよび仮想市場モデルは，セルが欠損であるかどうかだけが算定に影響する．そこで 0, 1 を値とする n 行 m 列の行列 X と，各属性の精神的損失レベル p，経済的損失レベル e，可用性 r，有用性 u，そして式 (4) における目的関数 v を定義するための t_1, t_2, t_3, t_4, c を入力として，局所解を出力とするプログラムとして，解法 3.5 と解法 4.7 を python で実装した．ただし解法 4.7 は入力としてさらに，何列まで全探索を可能とするかを指定する閾値 s も入力として必要とする．

実験の際には，現在の JO モデルおよび仮想市場モデルで採用されている底である $t_1 = 10$，$t_2 = 5$，$t_3 = 8$，$t_4 = 6$ を採用した．また，社会的責任度が 2 であるという想定の下，定数 c を 1 として設定した．また，各属性の精神的損失レベル p，経済的損失レベル e，可用性 r，有用性 u は，属性に対して $\{1, 2, 3\}$ のどれかの値を返すランダムな関数として与えた．その理由は，既存の取り組み [4], [1] で p, e, r, u が与えられている属性は限られているので，属性数が多い場合に対応できないからである．しかし，今回は二つの解法の性能比較実験をすることが目的であるため，ランダムに与えて問題ないと考えられる．

そこで，$n = 100$ 人の個人の $m = 30$ 個の属性を持つ行列に対して，解法 4.7 に対しては $s = 9$ 個の属性までは全探索可能であると設定したところ，解法 3.5 はおよそ 1.4 秒，解法 4.7 はおよそ 1.25 秒で，局所解を探索した．これらはほぼ同じ時間なので，同程度の計算量と推測できる．

5.2 実験

$r \in \{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$ に対して，欠損値の割合（すなわち 0 である割合）が $r\%$ である 0 または 1 を成分とする行列 X を 100 個ランダムに与える．それを $(X_1, \ldots, X_{100})$ とする．各 $i \in \{1, \ldots, 100\}$ に対し，X_i に対する解法 3.5 による近似解の仮想評価値を a_i，解法 4.7 による近似解の仮想評価値を b_i として，

$$\begin{aligned} \text{Win} &= \#\{i \in \{1, \ldots, 100\} \mid a_i < b_i\}, \\ \text{Lose} &= \#\{i \in \{1, \ldots, 100\} \mid a_i > b_i\}, \text{そして} \\ \text{Draw} &= \#\{i \in \{1, \ldots, 100\} \mid a_i = b_i\} \\ &\quad (= 100 - (\text{Win} + \text{Lose})) \end{aligned}$$

と定義する．Win は提案手法である解法 4.7 が，Lose は既存手法である解法 3.5 が，真に良い評価値を与えた回数である．Draw は同じ評価値を与えた回数である．

5.3 結果

実験結果を以下に表示する．

表 2 実験結果：試行回数は 100 回

欠損値の割合 $r\%$	Win	Lose	Draw
1%	25	0	75
2%	13	0	87
3%	22	0	78
4%	17	1	82
5%	22	0	78
10%	12	1	87
20%	9	5	86
30%	14	7	79
40%	16	6	78
50%	6	22	72

5.4 考察

今回の実験によって，$r \in \{1, 2, 3, 4, 5\}$ の場合は 2 割程度の確率で提案方式の方が真に良い値を与えることを観察できた．さらに，$r \in \{10, 20, 30, 40, 50\}$ について，$r \neq 50$ の場合は提案手法の方がより良い結果を与えることを観察できた．これらは，提案した本手法が欠損が少ない場合に有効なアルゴリズムであるという期待を，支持する実験結果となった．また $r = 4$ のとき Lose が 1 であること，$r \in \{1, 2, 3, 4, 5\}$ が増えるに応じて Win が減っていくわけではない原因は，既存手法である解法 3.5 で初期状態をランダムに取っているためであろうと推察している．

6. おわりに

本稿の主な背景は，表形式のパーソナルデータの価値とリスクを算定するモデルが与えられた場合，その差額「価値」−「リスク」として評価値を設定したときに，「どの属性を削除すればこの評価値が最大となるか」という問題を，全探索以外の手法で解くことにあった．JO モデルは [2] などですでに採用されているモデルであり，仮想市場モデルもヒアリングを通じた妥当性の確認を行っているモデルであるから，その差額を評価値として設定した．そこでこの評価値を最大化する属性削除問題 (= 問題 3.2) を提出し，その近似解を与える近似アルゴリズムである局所探索法 (= 解法 3.5) を紹介した．本稿では，「仮想市場モデルにおける本人特定容易度が常に 1 である」という設定の下で，本問題を問題 4.1 としてより厳密に定式化し，補助的な問題 (= 問題 4.3) とその解法 (= 解法 4.5) を用いることで，列削除問題に対する新しい発見法的解法アルゴリズムを提案した (= 解法 4.7)．さらに，局所探索法と今回提案したアルゴリズムの比較実験を行い，表の欠損が比較的少ない場合に，今回提案したアルゴリズムの方がより良い結果を与えていることを観察した．

将来的な課題としては，解法 4.7 の精度評価がある．特に，表データの $r\%$ がランダムに欠損したとき，解法 4.7 が最適でない確率を求める方法が課題として残っている．また，問題 4.1 は，ユークリッド空間上の区分的線形関数の 0-1 最適化問題とも定式化できる．こういった問題へのソルバーに関する知見に，関数 v (式 (4)) の形を活かすことで，効率のよい最適解を求めるアルゴリズムはないかという期待もあるが，現在は課題として残っている．

参考文献

[1] 一般財団法人日本情報経済社会推進協会：「匿名化技術等を活用した大規模なパーソナル情報の活用に関する調査研究」報告書 平成 23 年度次世代高信頼・省エネ型 IT 基盤技術開発・実証事業，日本情報経済社会推進協会 (2012).

[2] 一般社団法人日本サイバーセキュリティ・イノベーション委員会（JCIC）：「取締役会で議論するためのサイバーリスクの数値化モデル」(2018)，https://www.j-cic.com/pdf/report/QuantifyingCyberRiskSurvey-20180919(JP).pdf (参照 2019-08-22).

[3] 情報信託機能の認定スキームの在り方に関する検討会：「情報信託機能の認定に係る指針 ver1.0」，経済産業省 (2018)，https://www.meti.go.jp/press/2018/06/20180626002/20180626002-2.pdf (参照 2019-08-22).

[4] 特定非営利活動法人日本ネットワークセキュリティ協会セキュリティ被害調査ワーキンググループ：「２００３年度情報セキュリティインシデントに関する調査報告書」，特定非営利活動法人日本ネットワークセキュリティ協会 (2004)，http://www.jnsa.org/houkoku2003/incident_survey2.pdf (参照 2019-08-22).

[5] 特定非営利活動法人日本ネットワークセキュリティ協会セキュリティ被害調査ワーキンググループ：「情報セキュリティインシデントに関する調査報告書〜発生確率編〜」，特定非営利活動法人日本ネットワークセキュリティ協会 (2011)，https://www.jnsa.org/result/incident/data/2010incident_survey_probability.pdf (参照 2019-08-22).

[6] 山本 匡: 「情報漏えい被害の現状」, Network Security Forum 2004 講演資料 (2004), https://www.jnsa.org/nsf2004/kouen/A3.pdf (参照 2019-08-22).