行列分解を利用した 確率的 k-匿名性を満たす高次元データ公開法

長谷川 聡[1]　菊池 亮[1]　正木 彰伍[1]　濱田 浩気[1]

概要：本研究では, 購買データや映画の評価値データといった, 高次元なデータを匿名化する方法を提案する. 高次元データは, 次元の呪いにより有用性を高く保ったまま k-匿名化することが難しい. そのため従来の多くの研究では k-匿名性で考える攻撃者の背景知識を制限することで, 有用性を保ったデータを生成する方法が提案されてきた. しかしながら, これらの手法は安全性を下げることでしか次元の呪いを解決することができておらず, 従来の k-匿名性と同等の安全性を保った有用性の高いデータを生成することが望まれる. 本研究では, 高次元データをなるべく元のデータの特徴を保ちつつ低次元データに変換し, 低次元データに対して匿名化することで, 次元の呪いを回避しつつ有用性高い k-匿名性を満たすデータの生成を実現する. より具体的には, 購買データや評価値データなどを低次元データに変換する方法として知られている行列分解法と匿名化手法を組み合わせることで, 有用性の高い匿名化データを生成可能にする. 人工データを用いた数値実験により, 提案手法が, 高次元データでも有用性高く匿名化できることを確認したことを報告する.

キーワード：k-匿名化, Pk-匿名化, 行列分解, 高次元データ, 次元の呪い

Satoshi Hasegawa[1]　Ryo Kikuchi[1]　Shogo Masaki[1]　Koki Hamada[1]

1. はじめに

近年の人工知能やデータマイニング技術の発達により, 購買履歴や位置データといった個人に紐づく情報 (パーソナル情報) の二次利用が注目を浴びている. しかしながらこのようなパーソナル情報は, 個人特定可能な情報が含まれており, 他社に提供を行うとプライバシを侵害するリスクが生じる恐れがある. こうしたプライバシ侵害のリスクを低減し, データの第三者提供を可能にする方法として, 匿名化技術が研究されている [3], [5], [8], [9], [11], [17].

匿名化は, データベース中に含まれる個人データを加工し, 個人特定を困難にすることでプライバシ保護する技術をいう. 匿名化したデータがどの程度プライバシを保護しているかを示す代表的な指標として, k-匿名性 [11] が提案されている. k-匿名性とは, 「データベース中に同じ準識別子 (Quasi Identifier) を持つレコードが少なくとも k 個以上存在する」ことで個人のプライバシを保護する指標をいう. k-匿名性は, k 人未満に個人を絞り切れないという直感的にわかりやすい指標であるゆえ, 広く用いられている.

[1] NTT セキュアプラットフォーム研究所 180-8585 東京都武蔵野市緑町 3-9-11

	りんご	みかん	ばなな	・・・
山田太郎	1	0	0	・・・
佐藤花子	1	3	0	・・・
三田次郎	0	5	2	・・・
・・・	・・・	・・・	・・・	・・・

図 1　購買履歴データの例

本研究では, データの二次利用対象として注目をされている購買履歴データなどといったトランザクションデータに着目する. 購買履歴データの場合, 誰が何を何個買ったかが記されており, 購入対象となる商品そのものを準識別子として考えることが多い [12](図 1 参照). そのため購買履歴データは, 商品数分の次元数のデータと考えられ, 非常に高次元なデータとなる. よって, 高次元データに対し k-匿名性を満たす加工が必要となる.

1.1 高次元データに対する k-匿名化

k-匿名性を満たす匿名化の方法として, 属性の一般化による手法, クラスタリングによる手法が知られている [3], [8], [9]. これらの手法はいずれも高次元データで

は有用性の高い匿名化データを作ることが困難であることが示されている [1]. 高次元データの場合,　データベース中のレコード間の距離が急激に大きくなってしまうため, 距離の近いレコード同士をまとめるクラスタリングや一般化が途端に困難となることが主な理由としてあげられる (高次元データで k-匿名性を満たすことが困難になることは, 次元の呪いと呼ばれたりしている [1]).

高次元データで k-匿名性を満たすことが困難であることから, k-匿名性の仮定を緩めた匿名性指標が提案されている [4], [12]. 例えばトランザクションデータに対しては, k^m-匿名性が提案されている. これは, 攻撃者が持つ背景知識が高々 m 個のアイテムまでと制限した場合 (すなわち準識別子の数を制限する) に, 少なくとも k 個以上同じレコードが存在することを保証するプライバシ保護指標である [12]. この指標は k-匿名性を弱めた指標であり, m が全アイテム数の場合, 既存の k-匿名性と同等となる. 他にも時系列なデータに対し, LKC-privacy という指標 [4] などが提案されている. これは, 攻撃者の背景知識が L 系列までと制限することにより, 有用性の高い匿名化データを生成している. これらの手法は, 攻撃者の能力を明示的に示す必要があり, 攻撃者の能力が未知の場合や, 実運用を考えた場合は, m や L といったパラメータの設定が難しいことが課題である.

他に, データを一定の数の属性ごとに分割することにより, 次元の呪いを回避しながら匿名化する方法も考えられる [13]. しかしながらこの場合は, 属性ごとに分割されているために, 全属性を考慮した分析が行えない問題が発生してしまう. 仮に分割されたテーブル同士の結合を行おうとしても, 結合に必要な情報を持たないことから, 本来属性間が持つべき相関がないデータが出来上がってしまう. また, ある一定の属性ごとに独立にテーブルを分割できれば, 分割テーブル同士の結合も可能であるとも考えられるが, 購買データのような商品間に関係のあるようなデータセットは属性ごとに独立に分割することは困難であると考えられる.

1.2 貢献

本研究では, 攻撃者の能力を制限せず, 高次元データにおいても k-匿名性を満たす有用性の高いデータ生成方法を提案する. 高次元データにおいても k-匿名性を満たす有用性の高いデータを生成するためには, 次元の呪いを克服する必要がある. 次元の呪いを克服するアイディアとして, 高次元データをなるべく特徴を保持しながら低次元データに変換し, 低次元データに対し匿名化を行うことが考えられる. しかしながら, 元のデータは高次元データであるゆえ, 低次元データを再び高次元データに変換する必要が生じる. また, 低次元データが k-匿名性を満たしていたとしても, 高次元データに再変換した際に k-匿名性を満たすかどうかは自明ではない.

そこで, 高次元データへの再変換および k-匿名性を満たすことができる方法として, 購買データの推薦アルゴリズムなどで用いられている行列分解法に着目する [7], [10], [14]. 偶然にも購買データの推薦アルゴリズムとして用いられている行列分解を用いることで, 「高次元データ → 低次元データ → 匿名化された低次元データ → 匿名化された高次元データ」を生成することが可能となる.

また, 購買データは, 購入の有無を表す $\{0, 1\}$ データである可能性もある. そういったデータに対し, 一般化処理やクラスタリングを施すと意味のない値となってしまう (例えば一般化の場合は, 「0 また 1」という値となってしまう. またクラスタリングの場合は, 平均値を用いると, 小数が生じてしまう). そこで, 本研究ではランダム化による k-匿名性を満たす加工方法 (維持置換撹乱 [5], [15], [17]) により, この問題を解決する.

本論文の貢献は大きく 2 つある.

(1) 行列分解法を用いた次元の呪いを回避する匿名化方法の提案

(2) 提案法が k-匿名性を満たすことの保証

貢献 1 は, 行列分解法を用いて, 高次元データを低次元データに変換し, 低次元データに対し匿名化処理を施すことで, 次元の呪いを克服する. 貢献 2 は, 低次元データに対し k-匿名性を満たす匿名化処理を施したデータを, 再び高次元データに変換した際に, 高次元データにおいても k-匿名性を満たすことを示す.

本論文の構成を示す. 2 章では準備として, 記法の定義やランダム化による匿名化方法, 行列分解法を示す. 3 章では, 提案方法を示す. 4 章では, 実験結果を示し, 5 章では, まとめを述べる.

2. 準備

2.1 記法

本研究で扱うトランザクションデータの記法を定義する. トランザクションデータは, ユーザとアイテムの関係を表すデータであることから, 関係データと呼ばれており, 以後関係データと呼ぶこととする. また, 小文字の bold 体は列ベクトルを表し, 大文字の bold 体は行列を表す.

元データのレコード数 (ユーザ数) を N, アイテム数を M とする. i 番目のユーザと j 番目のアイテムとの関係の強さを x_{ij} で表す. 例えば, i 番目のユーザが j 番目のアイテムの購入数を表す場合は, $x_{ij} \in \mathbb{R}_+$ であり, また購入の有無を表す場合は, $x_{ij} \in \{\mathrm{TRUE}, \mathrm{FALSE}\}$ または $x_{ij} \in \{1, 0\}$ の 2 値のカテゴリ値で表す. j 番目のアイテムの取りうる値の集合を A_j と表し, カテゴリもしくは数値のどちらかを取るものとする. そして全アイテムの取りうる値の集合を $A = A_1 \times \cdots \times A_M$ とする. i 番目のユーザの全アイテムとの関係を表すベクトル $\mathbf{x}_i$ を, $\mathbf{x}_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{iM})^T \in A$ と表し, 全ユーザの全アイテムとの関係を表す行列 $\mathbf{X}$ を,

$$\begin{array}{c} \\ \text{ユーザ} \end{array} \begin{array}{c} \text{アイテム} \\ \begin{pmatrix} 0 & 2 & 0 & 3 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 4 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ \text{ユーザ} \end{array} \begin{array}{c} \text{アイテム} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

(a) 購買数を表す場合　　(b) 購買の有無を表す場合

図 2　購買データを模した関係データの例

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

と表す.

図 2 に関係データの例を示す. 図 2(a) は各ユーザが各アイテムを購入した数を表しており, 図 2(b) は各ユーザが各アイテムを購入したかどうかを表している. 例えば, 2 番目のユーザがそれぞれのアイテムを購入した数は $\mathbf{x}_2 = (1, 0, 2, 0)^T$ であり, 購入の有無は, $\mathbf{x}_2 = (1, 0, 1, 0)^T$ と表される.

2.2　ランダム化による k-匿名化 (撹乱再構築法)

ランダム化処理によって k-匿名性を満たす (正確には, k-匿名性を確率的な指標に拡張した Pk-匿名性を満たす) 方法として, 撹乱再構築法がある.

撹乱再構築法とは, 元データを条件付き確率 $P_{Y|X}$ に従いランダムに変更を加えることでデータを秘匿化し (撹乱と呼ぶ), 秘匿化されたデータから統計値を得る (再構築と呼ぶ) ことにより, プライバシを保護したまま統計分析を行う手法のことである.

既存の撹乱手法として, カテゴリ属性を撹乱する維持置換撹乱 [17], 数値属性を撹乱する方法として有界ラプラスノイズ [18] があり, 撹乱されたデータから元データの確率密度関数を推定する方法として逐次ベイズ法 [16], [18] が提案されている.

2.2.1　撹乱

維持置換撹乱

カテゴリ属性を撹乱する方法として, 維持置換撹乱が提案されている [2]. 維持置換撹乱とは, 維持確率 ρ で属性値を維持し, $1-\rho$ の確率で属性値をランダムに変更することで, データを秘匿化する処理である. あるカテゴリ属性 A_j の属性値 $v \in A_j$ が $v' \in A_j$ に変わる確率 $p_{y|x}^{A_j}(v'|v)$ は, 維持確率 ρ_j により,

$$p_{y|x}^{A_j}(v'|v) = \begin{cases} \rho_j + \frac{1-\rho_j}{|A_j|} & (v' = v) \\ \frac{1-\rho_j}{|A_j|} & (v \neq v') \end{cases} \tag{1}$$

と表すことができる.

有界ノイズ付与

数値属性を撹乱する方法として, 有界ラプラスノイズ付与や有界ガウスノイズ付与が提案されている [18]. 有界ラプラス分布とは, ラプラス分布の上限と下限が定まっている分布 (有界ラプラス分布) のことであり, この有界ラプラス分布に従う乱数を付与することで, データを秘匿化する. ある数値属性 A_j (値域が $[a_j, b_j], a_j \in \mathbb{R}, b_j \in \mathbb{R}$) の属性値 v が v' に変わる確率密度は, 有界ラプラス分布のパラメータ ϕ_j により,

$$P_{Y|X}^{A_j}(v'|v) = \frac{1}{\gamma_j(v)} \frac{1}{2\phi_j} \exp(-\frac{|v-v'|}{\phi_j}) \tag{2}$$

となる. ここで $\gamma_j(v)$ は以下である.

$$\gamma_j(v) = \int_{a_j - v}^{b_j - v} \frac{1}{2\phi_j} \exp(-\frac{|z|}{\phi_j}) dz \tag{3}$$

詳細ついては, 文献 [18] を参照されたい.

Pk-匿名性を満たすパラメータ決定法

ρ_j および ϕ_j は, 「撹乱後のテーブルのある人のレコードを $1/k$ 以上に確信することができない」(Pk-匿名性) を満たすようにする [5], [17], [18]. カテゴリ属性を $A_1, ..., A_l$, 数値属性を $A_{l+1}, ..., A_M$ とすると,

$$k = 1 + (N-1)\left(\prod_{1 \leq j \leq l} \left(\frac{1-\rho_j}{1 + (|A_j| - 1)\rho_j} \right)^2 \right) \prod_{l+1 \leq j \leq M} \exp\left(-2\frac{|b_j - a_j|}{\phi_j} \right) \tag{4}$$

が成立するよう ρ_j, ϕ_j を決めることにより, 維持置換撹乱および有界ラプラスノイズ付与による撹乱データは, Pk-匿名性を満たす.

上記のパラメータ決定方法は, 匿名化対象となるデータの中身に依存せず決定可能な方法であるゆえ, Pk-匿名性を満たすために強いノイズを加えている可能性がある. より弱いノイズで Pk-匿名性を達成する方法として, データのカテゴリ属性のみのデータに対し, データの中身に依存して維持確率 ρ_j を決定する方法が提案されている. 詳細については, 文献 [15] を参照されたい.

2.2.2　再構築

逐次ベイズ法

撹乱されたデータから元データの統計量を推定する方法として, 撹乱方法である $p_{y|x}$ および撹乱されたデータの度数 $h_y(\mathbf{y})$ から, 元データの確率密度関数 p_x を推定する方法が提案されている [16]. この方法は以下の対数尤度関数を最大化する (最尤推定) ことにより, p_x を推定する.

$$\arg\max_{\hat{p}_x} \sum_{m=1}^{|A|} h_y(\mathbf{y}_m) \log\left(\sum_{n=1}^{|A|} p_{y|x}(\mathbf{y}_m|\mathbf{x}_n) p_x(\mathbf{x}_n) \right) \quad \text{subject to} \sum_{n=1}^{|A|} p_x(\mathbf{x}_n) = 1, p_x(\mathbf{x}_n) \geq 0 \tag{5}$$

Agrawal ら [2] や五十嵐ら [17] は, $p_{y|x}$ に維持置換撹乱, h_y

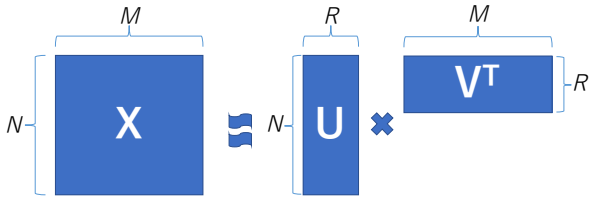

図 3 行列分解のイメージ. **X** を **U** と **V** の 2 つの行列の積で近似している.

として撹乱されたデータの度数分布, p_x として多次元ヒストグラムを仮定した再構築アルゴリズムを提案している. また式 (5) の解法として, EM アルゴリズムを適用した逐次ベイズ法を提案している [2].

五十嵐らは, 数値属性の撹乱データでも再構築ができるよう, $p_{y|x}$ として有界ラプラスノイズ付与を用いて逐次ベイズ法を適用する手法を提案している [18]. その際, 式 (2) を予め任意の区間で積分し, 確率密度から確率の値へ変換することで, 遷移確率行列を構成している.

2.3 行列分解

行列分解とは, 行列 $\mathbf{X} \in \mathbb{R}^{N\times M}$ を行列 $\mathbf{U} \in \mathbb{R}^{N\times R}$ と行列 $\mathbf{V} \in \mathbb{R}^{M\times R}$ の転置の 2 つの行列の積 (ただし, $R < N, M$) で近似することをいう (行列のランク R 近似ともいう, 図 3 参照).

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T \tag{6}$$

誤差が少なく近似できる場合は, 元の行列 **X** が低ランクであることが仮定できる時に限られる. しかし, 行列分解は, 購買データやレビューデータの推薦アルゴリズムとしても用いられている実績があり [6], 本研究においても誤差が少なく近似できることが期待できる.

行列を分解することにより, 様々な効果が得られる. 1 つは行列の持つ情報の量の削減である. 分解対象となる行列の **X** の要素数が NM であることに対し, **U** は NR, **V** は MR であるゆえ, $NR + MR$ の要素のみで $N \times M$ 行列を表現しているとわかる.

また, 次元圧縮効果もある. 分解した行列 **U** の各行ベクトル $\mathbf{u}_i$ を見た際, 元の行列 **X** の各行である M 次元のベクトル $\mathbf{x}_i$ を, R 次元のベクトル $\mathbf{u}_i$ に次元圧縮していると見ることができる. また, $\mathbf{u}_i$ と **V** との積により, R 次元に圧縮したベクトルを M 次元のベクトルに戻すことができる. 次元圧縮だけでなく, 元の次元数に戻すための情報も保持している点が, 単なる次元圧縮の方法と異なる.

3. 提案手法

本研究では, 高次元データの匿名化で問題となっていた次元の呪いによる有用性の低下を, 次元圧縮した低次元データに対してのみ匿名化すれば良いスキームを作ることで解決する.

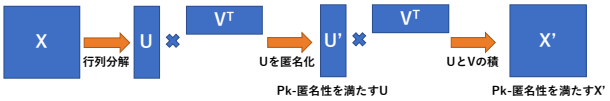

図 4 提案法の流れ

以下, 提案方式の概要を説明する (図 4 参照).

(1) 行列分解により, 元データ $\mathbf{X} \in \mathbb{R}^{N\times M}$ を行列 $\mathbf{U} \in \mathbb{R}^{N\times R}$, 行列 $\mathbf{V} \in \mathbb{R}^{M\times R}$ に分解する.

(2) 行列 **U** に対し *Pk*-匿名性を満たす撹乱処理 (維持置換撹乱や有界ノイズ付与など) を施した $\mathbf{U}'$ を得る ($\mathbf{U}'$ は *Pk*-匿名性を満たす).

(3) 撹乱済みの $\mathbf{U}'$ と **V** との積 $\mathbf{U}'\mathbf{V}^T$ により, $N \times M$ 行列 $\mathbf{X}'$ を得る. 得られた行列 $\mathbf{X}'$ は *Pk*-匿名性を満たしている.

提案方式は通常の匿名化よりも少ない撹乱度合いで匿名化が可能である. 従来どおりに **X** を *Pk*-匿名化する場合, 属性数 M で式 (4) を満たすようにパラメータ ρ_j や ϕ_j を求める必要がある. それに対し提案方式は, 属性数が $R(< M)$ な行列 **U** に対して式 (4) を評価するため, 撹乱度合いが少なく匿名化が可能である. ただし, R が小さければ小さいほど **X** の近似の誤差が大きくなるゆえ, 匿名化によるノイズと行列分解の近似とのバランスを取りながら本提案手法を用いる必要がある.

提案方式は **V** に対しては匿名化処理を加えていないが, $\mathbf{X}'$ は *Pk*- 匿名性を満たす. 行列分解を用いた匿名化に関し, 以下の定理を示す.

定理 1 行列 $\mathbf{X} \in \mathbb{R}^{N\times M}$ が行列 $\mathbf{U} \in \mathbb{R}^{N\times R}$, $\mathbf{V} \in \mathbb{R}^{M\times R}$ の行列の積で近似できるとする. このとき, **U** に対し *Pk*-匿名性を満たすランダム化処理 (もしくは k-匿名性を満たす一般化処理) を施した場合, 匿名化された $\mathbf{U}'$ を用いた, $\mathbf{X}' = \mathbf{U}'\mathbf{V}^T$ となる $\mathbf{X}'$ も *Pk*-匿名性 (k-匿名性) を満たす.

Pk-匿名性を満たす証明は付録で示す. k-匿名性を満たすことは, 容易に示すことができる. $\mathbf{X}'$ の各レコード $\mathbf{x}'_i$ は, $\mathbf{x}'_i = \mathbf{V}\mathbf{u}'_i$ である. すなわち $\mathbf{X}'$ の任意のレコード $\mathbf{x}'_i$ は, 対応する $\mathbf{u}'_i$ と共通の **V** との積で求まることから, **U** が k-匿名性を満たしていれば, $\mathbf{X}'$ は k-匿名性を満たす.

3.1 節では, **X** が 2 値行列の場合, 3.2 節では非負値かつ疎な行列の場合について, それぞれ有用性高く匿名化する具体的な方式を述べる.

3.1 X が 2 値行列の場合

X が 2 値行列, すなわち図 2(b) の場合の提案法を示す. 2 値行列を行列分解する方法は, binary matrix factorization[14] や boolean matrix factorization[10] などと呼ばれており, それぞれアルゴリズムが提案されている. 2 値行列分解は, 2 値行列 $\mathbf{X} \in \{0,1\}^{N\times M}$ を 2 値行列 $\mathbf{U} \in \{0,1\}^{N\times R}, \mathbf{V} \in \{0,1\}^{M\times R}$ に分解することをいう.

$$\mathbf{X}_{\{0,1\}} \approx \mathbf{U}_{\{0,1\}}\mathbf{V}^T_{\{0,1\}} \tag{7}$$

Algorithm 1 2 値行列の場合の行列分解を利用した *Pk*-匿名化法

Require: $\mathbf{X}_{\{0,1\}}, k, R$
Ensure: $\mathbf{X}'_{\{0,1\}}$
Factorize: 2 値行列分解により, $\mathbf{X}_{\{0,1\}} \approx \mathbf{U}_{\{0,1\}}\mathbf{V}^T_{\{0,1\}}$ となる $\mathbf{U}_{\{0,1\}}$, $\mathbf{V}_{\{0,1\}}$ を得る.
Randomize: $\mathbf{U}_{\{0,1\}}$ に対し, 式 (4) で維持確率 ρ を評価. $\mathbf{U}'_{\{0,1\}}$ を維持置換撹乱し, *Pk*-匿名性を満たす行列 $\mathbf{U}_{\{0,1\}}$ を得る.
Assemble: $\mathbf{U}'_{\{0,1\}}$ および $\mathbf{V}_{\{0,1\}}$ の行列積により, $\mathbf{X}'_{\{0,1\}} = \mathbf{U}'_{\{0,1\}}\mathbf{V}^T_{\{0,1\}}$ を得る.

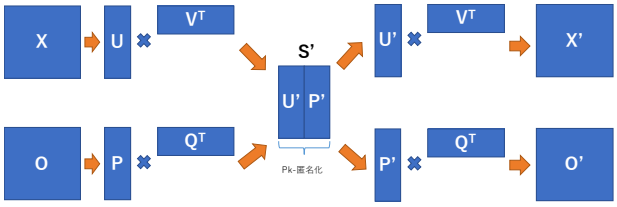

図 5 非負値かつ疎な行列の場合の提案法の流れ

$\mathbf{U}$ の要素は $\{0,1\}$ の 2 値しかとらない. それゆえ $\mathbf{U}$ を *Pk*-匿名性を満たすよう撹乱を行う場合, $\mathbf{U}$ の要素をカテゴリ値として捉え, 維持置換撹乱を用いることで, 有用性高く匿名化できることが期待される.

2 値行列の場合の提案方式をアルゴリズム 1 に示す.

3.2 X が非負値かつ疎な行列の場合

$\mathbf{X}$ が非負値かつ疎な行列, すなわち図 2(a) の場合の提案法を示す.

非負値な行列を行列分解する方法は, non-negative matrix factorization などと呼ばれている [7]. 非負値行列分解は, 非負値行列 $\mathbf{X} \in \mathbb{R}_+^{N\times M}$ を非負値行列 $\mathbf{U} \in \mathbb{R}_+^{N\times R}, \mathbf{V} \in \mathbb{R}_+^{M\times R}$ に分解することをいう.

$$\mathbf{X}_+ \approx \mathbf{U}_+\mathbf{V}_+^T \tag{8}$$

特に疎な行列を対象とする場合は, 疎な要素のみに注目し, 行列分解した結果が元行列の疎要素のみと近くなるように分解をしている. 非負値かつ疎な行列を $\mathbf{X}_+$ とする. また $\mathbf{O}^{N\times M}_{\{0,1\}}$ の i 行 j 列の要素 o_{ij} を

$$o_{ij} := \begin{cases} 1 & \text{if } x_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

と定義する. 疎な場合の非負値行列分解は以下の最適化問題を解く ($*$ は行列の要素積を表す).

$$\arg\min_{\mathbf{U}_+,\mathbf{V}_+} \|\mathbf{O} * (\mathbf{X}_+ - \mathbf{U}_+\mathbf{V}_+^T)\|^2_{Fro} \tag{10}$$

ここで, $\|\cdot\|_{Fro}$ はフロベニウスノルムを表す.

この場合, 非ゼロ要素の誤差のみ少なくよう行列分解を行うため, 全要素を考慮した行列分解より, 小さいランク R で誤差が少なく行列分解が可能となる. ただし誤差が少ないのはあくまで非ゼロ要素のみであり, $\mathbf{U}_+\mathbf{V}_+^T$ の結果は疎でなくなってしまう. それゆえ非負値かつ疎な行列を, 行列分解を用いて匿名化を行う場合は, 非ゼロ要素を表す情報が別途必要となる.

Algorithm 2 非負値かつ疎な行列の場合の行列分解を利用した *Pk*-匿名化法

Require: $\mathbf{X}, k, R$
Ensure: $\mathbf{X}'$
Factorize: 非負値行列分解により, $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$ となる $\mathbf{U}, \mathbf{V}$ を得る. また, バイナリ行列分解により, $\mathbf{O} \approx \mathbf{P}\mathbf{Q}^T$ となる $\mathbf{P}, \mathbf{Q}$ を得る.
Randomize: $\mathbf{U}$ と $\mathbf{P}$ を水平結合した $\mathbf{S}$ に対し, 式 (4) で維持確率 ρ およびパラメータ ϕ を評価. $\mathbf{U}$ に対し有界ノイズ付与を適用し, $\mathbf{P}$ に対し維持置換撹乱を適用し, *Pk*-匿名性を満たす行列 $\mathbf{U}', \mathbf{P}'$ を得る.
Assemble: $\mathbf{U}'$ および $\mathbf{V}$ の行列積により, $\mathbf{X}' = \mathbf{U}'\mathbf{V}^T$ を得る. また $\mathbf{P}'$ および $\mathbf{Q}$ の行列積により, $\mathbf{O}' = \mathbf{P}'\mathbf{Q}^T$ を得る.
Masking: $\mathbf{O}'$ で, $\mathbf{X}'$ をマスクする.

表 1 情報損失の評価

	ランク 3	ランク 5	ランク 10	ランク 20
提案法	0.0102	0.0115	0.0162	0.0932
k-匿名化	0.0176	0.0183	0.0193	0.0265
Pk-匿名化	0.5633	0.5553	0.5631	0.5639

そこで, 3.1 節で提案した, 2 値行列の場合の匿名化方法と組み合わせることで, 疎性を保った匿名化データを生成する方法を提案する. $\mathbf{O}_{\{0,1\}}$ を, 2 値行列分解した結果を,

$$\mathbf{O}_{\{0,1\}} \approx \mathbf{P}_{\{0,1\}}\mathbf{Q}^T_{\{0,1\}}, \tag{11}$$

とする. そして, $\mathbf{U}_+ \in \mathbb{R}_+^{N\times R}$ と $\mathbf{P} \in \mathbb{R}_{0,1}^{N\times R}$ とを水平結合した新たな行列 $\mathbf{S} \in \mathbb{R}^{N\times 2R}$ に対し, 匿名化処理を行う. その際, $\mathbf{U}_+$ の要素は正の実数であるゆえ有界ノイズ付与を, $\mathbf{P}_{\{0,1\}}$ の要素は 0,1 であるゆえ 3.1 節の方法と同様に維持置換撹乱を適用する. 匿名化された $\mathbf{S}'$ を $\mathbf{U}'_+$ および $\mathbf{P}'$ に分割し, それぞれ $\mathbf{V}_+$ および $\mathbf{Q}$ を掛け合わせることで, 元の行列に戻す. そして $\mathbf{O}'$ で $\mathbf{X}'$ をマスクすることにより, *Pk*-匿名化されたデータを得る.

非負値かつ疎な行列の場合の提案方式の流れをアルゴリズム 2 に示す (図 5 にアルゴリズム 2 の概念図を示す).

4. 実験

人工的に作成したデータセットを用いた数値実験により, 提案法の有効性を示す.

4.1 実験設定

実験データとして, 行列の要素のうちおよそ 1%のみ 1 を持ち, かつ行列のランクがそれぞれ 3, 5, 10, 20 である 1000×1000 行列 $\mathbf{X} \in \{0,1\}^{1000\times 1000}$ を作成し, 作成した行列の 1% のランダムな要素に対しノイズ ($\{0,1\}$ を入れ替える) を加えた行列を用いる.

提案手法との比較対象として, microaggregation を用いた *k*-匿名性を満たす匿名化法, 従来の維持置換撹乱を用いた

Pk-匿名化手法との比較を行う. また, $k = 10$ として実験する.

評価指標として, 匿名化前のデータ $\mathbf{X}$ と匿名化後のデータ $\mathbf{X}'$ でどの程度誤差が生じているかを測るために, 式 (12) に定義した情報損失を用いる.

$$\text{loss} := \frac{1}{NM} \sum_{ij} |x_{ij} - x'_{ij}| \tag{12}$$

提案法は, 対象となる行列が 2 値であることから, アルゴリズム 1 を用いる. 2 値行列分解におけるランクは $R = \{2, ..., 20\}$ まで変化させ, 得られた匿名化データの情報損失最も小さい場合を提案手法の結果として採用する.

4.2 実験結果

実験結果を表 1 に示す. 従来の *Pk*-匿名化は, 高次元データにおいて情報損失がとても大きいことがわかる.

ランク 10 までのデータに関しては, 提案法が最も情報損失が少ないことがわかる. *Pk*-匿名化は, サンプル数に対し, 属性の値の取りうる組み合わせ数が少ないほうが, 有用性高くデータを生成できる特徴がある. 提案法において, $\mathbf{U}$ が 10 次元 (ランク 10 で行列分解した場合の $\mathbf{U}$) の場合は, 属性の値の取りうる組み合わせ数が最大でも $2^{10} = 1024$ 通りであり, 実際には十数通り程度の組み合わせしかなかったことから, サンプル数 1000 でも有用性高く *Pk*-匿名化が実施でき, このような結果が得られたと考えられる.

また, ランク 20 の場合に提案法の損失が増えた原因として, 匿名化に必要なサンプル数が少なかったことがあげられる. 具体的には, $\mathbf{U}$ が 20 次元の場合, 取りうる値の組み合わせの最大数が $2^{20} = 1048576$ であり, 実際にも数百通りの組み合わせがあったため, サンプル数 1000 では足りず, 匿名化がうまくできなかったと考えられる.

実験結果より, サンプル数が十分かつ行列の低ランク性が仮定できる場合, 提案法が有効であることがわかった.

5. おわりに

高次元データを *k*-匿名性を満たしながら公開する際の問題であった次元の呪いを改善する匿名化アルゴリズムを提案した. 提案方法は, 一度データを低次元に圧縮することで, 匿名化の際に生じる次元の呪いによる有用性の低下を回避しつつ, *k*-匿名性を満たした有用性の高いデータを作成することが期待できる. 人工的に作成したデータセットにおいては, 元のデータにある程度の低ランク性が仮定できる場合, 提案手法が有効に働くことが確認できた.

今後の課題は, 提案したアルゴリズム 2 の効果の確認および実データを用いた実験による提案手法の有効性の評価などがあげられる. また, *k*-匿名性を満たす匿名化手法との比較は行ったが, 攻撃者の背景知識を制限した k^m-匿名性を満たす匿名化手法などとは比較できておらず, 今後の課題である.

参考文献

[1] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pp. 901–909. VLDB Endowment, 2005.

[2] Rakesh Agrawal, Ramakrishnan Srikant, and Dilys Thomas. Privacy preserving olap. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 251–262, 2005.

[3] Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. Efficient k-anonymization using clustering techniques. In *Proceedings of the 12th International Conference on Database Systems for Advanced Applications*, pp. 188–200, 2007.

[4] Benjamin Fung, Ming Cao, Bipin C Desai, and Heng Xu. Privacy protection for rfid data. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pp. 1528–1535. ACM, 2009.

[5] Dai Ikarashi, Ryo Kikuchi, Koji Chida, and Katsumi Takahashi. *k* - anonymous microdata release via post randomization method. In *International Workshop on Security 2015*, pp. 225–241, 2015.

[6] Yehuda Koren, Robert Bell, Chris Volinsky, et al. Matrix factorization techniques for recommender systems. *Computer*, pp. 30–37, 2009.

[7] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

[8] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60, 2005.

[9] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, pp. 25–25, 2006.

[10] Siamak Ravanbakhsh, Barnabás Póczos, and Russell Greiner. Boolean matrix factorization and noisy completion via message passing. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 945–954, 2016.

[11] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570, 2002.

[12] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, Vol. 1, No. 1, pp. 115–125, 2008.

[13] Hessam Zakerzadeh, Charu C Aggarwal, and Ken Barker. Towards breaking the curse of dimensionality for high-dimensional privacy. In *SDM*, pp. 731–739. SIAM, 2014.

[14] Zhongyuan Zhang, Tao Li, Chris Ding, and Xiangsun Zhang. Binary matrix factorization with applications. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 391–400. IEEE, 2007.

[15] 菊池亮, 五十嵐大, 千田浩司, 濱田浩気. データ分布依存処理によって高い有用性を実現する確率的 k-匿名性. 暗号と情報セキュリティシンポジウム 2013, 2013.

[16] 五十嵐大, 千田浩司, 高橋克巳. 多値属性に適用可能な効率的プライバシー保護クロス集計. コンピュータセキュリティシンポジウム 2008, pp. 497–502, 2008.

[17] 五十嵐大, 千田浩司, 高橋克巳. k-匿名性の確率的指標への

拡張とその適用例. コンピュータセキュリティシンポジウム 2009, pp. 1–6, 2009.
[18] 五十嵐大, 長谷川聡, 納竜也, 菊池亮, 千田浩司. 数値属性に適用可能な, ランダム化により k-匿名性を保証するプライバシー保護クロス集計. コンピュータセキュリティシンポジウム 2012, pp. 639–646, 2012.

付　　録

A.1　定理 1 の証明

証明 1　$\mathbf{X}'$ の各レコードが $1/k$ 以上の確率で絞り込むことができないことを示すため,

(1) $\mathbf{U}$ と $\mathbf{U}'$ の識別問題 ($\mathbf{U}$ が与えられた時に $\mathbf{U}'$ が Pk-匿名性を満たす)
(2) $\mathbf{U}$ と $\mathbf{X}'$ の識別問題 ($\mathbf{U}$ が与えられた時に $\mathbf{X}'$ が Pk-匿名性を満たす)
(3) $\mathbf{X}$ と $\mathbf{X}'$ の識別問題 ($\mathbf{X}$ が与えられた時に $\mathbf{X}'$ が Pk-匿名性を満たす)

の三段階を踏んで $\mathbf{X}'$ の Pk-匿名性を示す.

U と U' の識別問題

$\mathbf{U}'$ は, $\mathbf{U}$ を *Pk*-匿名性を満たすように生成している. それゆえ, $\mathbf{U}$ と $\mathbf{U}'$ 間のレコードの識別問題を考えた場合, $1/k$ 以上の確率に絞り込めないことが保証されている.

U と X' の識別問題

$\mathbf{U}$ と $\mathbf{X}'$ との識別問題を考えるにあたり, まず $\mathbf{U}'$ と $\mathbf{X}'$ との関係を考える. $\mathbf{U}'$ の各レコード $\mathbf{u}'_i$ は, $\mathbf{X}'$ の各レコード $\mathbf{x}'_i$ と 1 対 1 対応している. もし, $\mathbf{U}$ と $\mathbf{X}'$ とのレコードの識別問題を考えた際に $1/k$ 以上の確率であるレコードを絞り込めた場合, $\mathbf{U}'$ と $\mathbf{X}'$ が 1 対 1 対応しているゆえ, $\mathbf{U}'$ のレコードも $1/k$ 以上の確率で絞り込めることを意味する. しかしながら, $\mathbf{U}'$ の各レコードは $1/k$ 以上の確率で絞り込めないようことが保証されている (*Pk*-匿名化を行っているゆえ). それゆえ, $\mathbf{U}$ と $\mathbf{X}'$ とのレコードの識別問題を考えた場合, $1/k$ 以上の確率でレコードを絞り込むことができない.

X と X' の識別問題

最後に $\mathbf{X}$ と $\mathbf{X}'$ との関係を考える. $\mathbf{U}$ と $\mathbf{X}$ の各レコードは 1 対 1 対応している. もし $\mathbf{X}$ と $\mathbf{U}'$ とのレコードの識別問題を考えた際に $1/k$ 以上の確率であるレコードを絞り込めた場合, $\mathbf{X}$ と $\mathbf{U}$ が 1 対 1 対応しているゆえ, $\mathbf{U}'$ のレコードも $1/k$ 以上の確率で絞り込めることを意味する. しかしながら, $\mathbf{U}'$ の各レコードは $1/k$ 以上の確率で絞り込めないようことが保証されている (*Pk*-匿名化を行っているゆえ). それゆえ, $\mathbf{X}$ と $\mathbf{X}'$ とのレコードの識別問題を考えた場合, $1/k$ 以上の確率でレコードを絞り込むことができない.　□