情報保護の統計モデル

星野　伸明
金沢大・経

2014 年 11 月 19 日

概要

- （ビッグ）データを社会で利用する際の問題の一つ：情報保護
- しかし情報保護研究の成果は十分活用されていないように見える。
  - 研究成果と社会的要請との関係が必ずしも明らかでない。
  - 機械学習の結果が解釈可能でないと使われにくいのと同じ。
- 法に明示される社会的要請を、技術的な問題へ直接に「翻訳」して解けばよかろう。
- 例として匿名性管理の問題を示す。

情報保護とは

- 情報保護はプライバシーを守るための手段である。
  - プライバシー概念：自己情報をコントロールする権利として理解する立場が有力。
  - 秘密の自己情報は知られないようにコントロールしたい。
- 特定個体の秘密をデータが暴露する (disclose) ということ：
  1. 「推測暴露」: レコード情報の主が分からなくても分かる状態。
     - 例）ある建物居住者は全員年収 2000 万円以上と分かる。
  2. 「識別暴露」: レコード情報の主が判明して分かる状態。
     - 例）{ 職業＝首相, 国籍＝日本, 病状＝ · · ·}
- 情報保護の技術的研究はこれら暴露概念の詳細な定義から。

推測暴露について

- 識別暴露よりも技術的に定義しやすい。
  - 基本的には、個体属性所与で秘密変数の条件付き分布の分散が小さければ「暴露」と考える。
- 推測暴露でない（分散が大きい）状態の技術的概念：
  1. 例）ℓ– 多様性 (diversity)：個体属性所与で秘密変数の観測値が ℓ 通り以上あること。
  2. 例）t-近接性 (closeness)：個体属性所与で秘密変数の条件付き分布が、非条件付き分布と近いこと（＝郡内分散が落ちない)。
  3. 例）差分 (differential) プライバシー：秘密変数の尤度関数が平らということ（次頁)。

ϵ-差分プライバシー（流行中）

- 元データ $\boldsymbol{X} = \boldsymbol{x}$ から、機構 μ により（ランダムノイズを乗せて）保護したデータ $\boldsymbol{Y}$ を生成する。
- 一要素（レコードやフィールド）だけ違う $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ について

$$\left| \max_{\boldsymbol{x}_1, \boldsymbol{x}_2} \frac{\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}_1)}{\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}_2)} \right| \leq \exp(\epsilon), \quad \forall \boldsymbol{y}, \tag{1}$$

  なら μ は「ϵ-差分プライバシー」を保証すると言う。
- 式 (1) で $\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}_1) / \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}_2)$ は母数 $(\boldsymbol{x})$ の尤度比。これが全域で 1 に近ければ尤度関数は平らになり、母数推定の精度は出ない。
  - 要素がレコードなら、ある個体が公開データに入っているか分からない。
  - 要素がフィールドなら、その変数の値が分からない。

推測暴露研究の問題 (私見)

- データの分析価値を著しく損なう。
  - 保護強度 (ℓ, t, ϵ) を緩和すると保護はピンぼけ。
- データ分析（秘密変数に興味）も社会的要請。
  - 社会的にプライバシーが優先だが。
- 技術的な情報保護概念が社会的要請とずれているのでは。
  - 保護強度の問題ではなく、秘密変数への態度が問題。
- データの分析価値が残るような情報保護概念を工夫しないと、技術的研究成果が社会で活用されない。
  - 秘密変数の統計的利用を許すような reasoning が必要。

曖昧な社会的要請の具体化

- 社会的要請に適合する（情報保護）概念を構成するには、法の参照から始めてはどうか。
  - 法は最低限の社会的要請を明示。
  - 法の文言は曖昧だが、更に曖昧な（プライバシー保護）概念を限定したもの。
- 法的には、個体識別を避けたい。
  - 例）個人情報の定義：「生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日その他の記述等により特定の個人を**識別**することができるもの（他の情報と容易に**照合**することができ、それにより特定の個人を識別することができることとなるものを含む。）」
- 識別暴露抑止の方が社会的要請に近い。

個体識別とは？

- 「個体が識別できる」のおおよその意味 : あるレコードが特定個体の情報と分かること。
- 「照合」は個体識別の近道として知られている。
- 公開情報と既知の「攻撃用（個体）情報」を照合して、一意に一致したレコードは同じ個体かもしれない。
  1. 公開個体群と既知の個体群が同じ時、一意に一致なら個体識別と言える。
  2. 公開個体群と既知の個体群が同じとは限らない時、一意に一致、
     かつ非公開群に同属性の個体が存在しない なら個体識別と言える。
  - 「非公開群」を (母集団 \ 公開標本) と考える。

- 例１）非公開群に同属性の個体が存在するので識別不可能：
  - 母集団：{(A さん, 男, 38 才,-),(B さん, 女, 42 才,-),(C さん, 男, 38 才,+)}
  - 公開標本：{(男, 38 才,+)}
  - 攻撃用情報:{(C さん, 男, 38 才)}
- 例２）非公開群に同属性の個体が存在しないので識別可能：
  - 母集団：{(D さん, 男, 48 才,-),(B さん, 女, 42 才,-),(C さん, 男, 38 才,+)}
  - 公開標本：{(男, 38 才,+)}
  - 攻撃用情報:{(C さん, 男, 38 才)}
- 例３）もし母集団が C さんだけなら非公開群は $\emptyset$ で識別可能。
  - 母集団：{(C さん, 男, 38 才,+)}
  - 公開標本：{(男, 38 才,+)}
  - 攻撃用情報:{(C さん, 男, 38 才)}

- 個体識別と言える条件を書き直すと「攻撃用情報と照合して一意に一致する標本の個体が母集団でも一意」。
  - 照合に使う既知の変数を「キー変数」または「擬似識別子」と呼ぶ。
    - 例 1-3 では、性別と年齢がキー変数。
- 従って、所与のキー変数について「標本一意」かつ「母集団一意」のレコードが個体識別の危険性が高いと考えられている。
- 危険なレコードを減らすには、キー変数を操作する。
  - 秘密変数はいじる必要がないのでデータの分析価値を保ちやすい。

- 例４）年齢の「再符号化」＝「一般化」により識別不可能：
  - 母集団：{(D さん, 男, 48 才,-),(B さん, 女, 42 才,-),(C さん, 男, 38 才,+)}
  - 公開標本：{(男, 30-50 才,+)} は、標本一意かつ母集団二意
  - 攻撃用情報:{(C さん, 男, 38 才)}
- 例５）性別の「削除」によりますます識別不可能：
  - 母集団：{(D さん, 男, 48 才,-),(B さん, 女, 42 才,-),(C さん, 男, 38 才,+)}
  - 公開標本：{(x, 30-50 才,+)} は、標本一意かつ母集団三意
  - 攻撃用情報:{(C さん, 男, 38 才)}

識別暴露の管理

- 標本中の母集団一意数を受容限界以下に管理したい。
  - 一意を全て消す k-匿名基準は過保護と思われる。
- 母集団一意数は通常は未知なので推定する。
- サンプリングの不確実性を扱うには統計モデルが便利。
  - 属性ベクトル $\boldsymbol{x}$ をインデクスとする確率変数 $F_{\boldsymbol{x}}$ が非負整数上の分布に従う。
  - 母集団で $F_{\boldsymbol{x}} = 1$ となる属性が母集団一意。

属性空間の分割

- 公開レコードの属性 $\boldsymbol{x}$ 同士に包含関係があると母集団一意は識別の条件にならない。
  - 例）公開レコード：$\{\boldsymbol{x}_1 =$(10-30 才, 男), $\boldsymbol{x}_2 =$(15-20 才, 男)$\}$、母集団で $F_{\boldsymbol{x}_1} = 1, F_{\boldsymbol{x}_2} = 1$ とする。しかし第二レコードは識別不可能。
- 通常は属性空間を分割した「セル」に個体が分布するような匿名化をする。
  - 例）官庁統計では全レコードに同一基準の「大域的」再符号化を施すのが基本。この場合、高次元の「クロス集計表」と公開データセットは同値となる。
- その場合、分割表の統計モデルが適切。

個票データの分割表表現

- 大域的再符号化のみ使われる場合、個票データセットと分割表は一対一対応。

Table 1: 個票データ例

Name	Sex	Age
x	M	-14
x	M	-14
x	M	15-64
x	M	65-
x	F	-14
x	F	15-64
x	F	15-64

Table 2: 2×3 分割表

Sex \ Age	-14	15-64	65-
F	1	2	0
M	2	1	1

- Table 1 と 2 は一対一対応

ビッグデータの匿名性

- ビッグデータの匿名化ならではの特性：**キー変数が多い＝分割表の次元 p が高い**
  - 例）ライフログや軌跡などは各記録時点それぞれキー変数。
- 次元の呪いが強いので、「疎な分割表」をモデル化したい。
  - 疎な分割表の統計モデルは漸近論 $n \to \infty$ に持ち込むのが普通。
    - 例）現代的な漸近論：$p/n \to const.$
    - 例）古典的な漸近論：$J/n \to const.$, ただし J は総セル数。
      - 各キー変数のカテゴリー数が c なら $J = c^p$
  - しかし母集団サイズ n 所与で度数 1 のセル数に興味があるので、普通の漸近論は必ずしも適切でない。
  - 以下では n を固定し $J \to \infty$ の極限をとる枠組み (Hoshino, 2012) を説明。

分割表についての記法

- セル総数（非確率変数） = J
- 第 j セルの個体数 = $F_{j,J}, j = 1, 2, \ldots, J.$

$$\boldsymbol{F}_J := (F_{1,J}, \ldots, F_{J,J})$$

- 個体数 i のセル数 = $S_{i,J}$:「寸法指標」、「頻度の頻度」

$$\boldsymbol{S}_J := (S_{1,J}, S_{2,J}, \ldots).$$

  - 一意数は $S_{1,J}$.
    - （スパース回帰で）$\boldsymbol{F}_J$ を予測するより $S_{1,J}$ の予測の方が楽なはず。
- 総個体数 = $N_J = \sum_{j=1}^{J} F_{j,J} = \sum_{i=1}^{\infty} iS_{i,J}$.

疎な分割表上の小数法則

- 単純に $J \to \infty$ とすればモデルが発散するので、$\mathsf{E}(N_J)$ が固定されるように基準化。
  - 二項分布の期待値 $np = \lambda$ を固定して $n \to \infty, p \to 0$ とした極限は平均 λ のポアソン分布（$\mathsf{Po}(\lambda)$）となる。これを「小数法則」と呼ぶ。
- 極限で度数が 0 のセル数は無限大なので、度数が 1 以上のセルの挙動、すなわち $\boldsymbol{S}_J$ の極限分布を用いる。
- n を固定するには $(\boldsymbol{S}_J | N_J)$ の条件付き分布とその極限を用いる。
- 以下の正準条件を課す：

$$0 < \mu < \infty, \quad q_i \geq 0, i \in \mathbb{N}, \quad \sum_{i=1}^{\infty} q_i = 1.$$

  - このとき $(q_i)_{i=1}^{\infty}$ は正の整数上の分布を与える。

命題 1 (Koopman, 1950) 非負整数上の分布に従う確率変数 $F_{j,J}, j \in [J]$, は互いに独立とする。

この時

$$\boldsymbol{S}_J \xrightarrow{d} \mathop{\perp}_{i=1}^{\infty} \mathsf{Po}(\mu q_i) \tag{2}$$

と以下の条件は同値である：

$$\lim_{J\to\infty} \max_j \mathsf{P}(F_{j,J} = i) = 0, \quad i \in \mathbb{N}, \tag{3}$$

かつ

$$\lim_{J\to\infty} \mathsf{E}(S_{i,J}) = \mu q_i, \quad i \in \mathbb{N}. \tag{4}$$

- 式 (3) は、度数が正となる確率が「一様にほとんど無視できる」ということ。

命題 2 (Wang and Ji, 1993) 非負整数上の分布に従う確率変数 $F_{j,J}, j \in [J]$, は互いに独立とする。この時 (2) と

$$N_J \xrightarrow{d} \mathtt{CP}(\mu, \boldsymbol{q})$$

は同値である。ただし $\mathtt{CP}(\mu, \boldsymbol{q})$ は以下の確率母関数で定義される「複合ポアソン分布」:

$$G(z) = \exp(\mu(g(z) - 1)), \quad \mu > 0,$$

$$g(z) = \sum_{i=1}^{\infty} z^i q_i.$$

- $g(z)$ は $(q_i)_{i=1}^{\infty}$ の確率母関数。
- $g(z) = z$ なら $G(z)$ はポアソン分布。

複合ポアソン分布の例

$G(z)$	$g(z)$
負の二項分布:	対数級数分布
対数正規＝ポアソン分布:	?
一般化逆ガウス＝ポアソン分布:	?
ポアソン・パスカル分布:	拡張負の二項分布
ラグランジュアン・ポアソン分布:	ボレル分布

命題 3 $N_J = n$ 所与で $\boldsymbol{S}_J$ の条件付き分布は、条件 (3) と (4) を満たす極限操作 $J \to \infty$ により以下の分布に収束する。

$$\mathsf{P}(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = \frac{n!\mu^u \prod_{i=1}^n {q_i}^{s_i} \frac{1}{s_i!}}{B_n(\mu x_1, \ldots, \mu x_n)}, \tag{5}$$

$$(s_1, \ldots, s_n) \in \mathcal{S}_{|n} := \{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i = 1, 2, \ldots, n, \sum_{i=1}^n i s_i = n\}.$$

ただし $x_i = i!\, q_i, u = \sum_{i=1}^n s_i$, そして基準化定数 $B_n(\mu x_1, \ldots, \mu x_n)$ は「全ベル多項式」である。

- 確率関数 (5) を持つ分布を「母数 $(\mu, \boldsymbol{q})$ の Limiting Conditional Compound Poisson (LCCP) 分布」と呼ぶ。
- $\mathcal{S}_{|n}$ 上の分布を自然数 n の「確率分割」と呼ぶ。

例） Ewens 分布 (Ewens, 1972)

- 母数 $(-\kappa \log(1-\theta)$, 対数級数分布$)$ の LCCP 分布は Ewens 分布：

$$\mathsf{P}(\boldsymbol{S}=\boldsymbol{s}|N=n)=\frac{\kappa^k n!}{\kappa(\kappa+1)\cdots(\kappa+n-1)}\prod_{i=1}^{n}\left(\frac{1}{i}\right)^{s_i}\frac{1}{s_i!}, \tag{6}$$

$$\boldsymbol{s}\in\mathcal{S}_{|n}.$$

- $B_n(\kappa 0!, \kappa 1!, \ldots, \kappa(n-1)!)=\kappa(\kappa+1)\cdots(\kappa+n-1)$.
- $F_{j,J}$ が負の二項分布なら、$\boldsymbol{F}_J|N_J$ は負の超幾何分布に従う。この分布は多項分布のセル確率がディリクレ分布に従うサンプリングと同じ（ノンパラベイズの典型)。その $J\to\infty$ の極限分布が Ewens 分布。
- Pitman-Yor の Chinese Restaurant Process (CRP) では $\alpha=0$ の場合のみ LCCP 分布。

LCCP 分布による母集団一意数評価

- 未知母数 μ の推定の挙動は良好。
  - 母数 $(\mu, \boldsymbol{q})$ の LCCP 分布は $\boldsymbol{q}$ 所与で指数族。そして母数 μ の十分統計量は $S_1 + \cdots + S_n$ である。
- 母集団サイズを n として S_1 の周辺分布や期待値は次頁以下のように評価できる。
- $n \to \infty$ なら $S_i \xrightarrow{d} \mathsf{Po}(\mu q_i)$.
  - LCCP 分布は $\boldsymbol{S} \sim \perp_{i=1}^{\infty} \mathsf{Po}(\mu q_i)$ を $N = n$ で条件付けしている。

同時階乗モメント

命題 4 $(S_1, \ldots, S_n)$ が母数 $(\mu, \boldsymbol{q})$ の *LCCP* 分布に従うとする。このとき $R := \sum_{i=1}^{n} i r_i \leq n$ を満たす全ての $r_1, \ldots, r_n \in \mathbb{N}_0$ について

$$E(\prod_{i=1}^{n} S_i^{(r_i)}) = \frac{B_{n-R}(\mu x_1, \ldots, \mu x_{n-R}) \mu^r n^{(R)}}{B_n(\mu x_1, \ldots, \mu x_n)} \prod_{i=1}^{n} (\frac{x_i}{i!})^{r_i}, \tag{7}$$

ただし $r = \sum_{i=1}^{n} r_i$ そして $n^{(R)} = n(n-1)\cdots(n-R+1)$.

階乗モメントの逆転公式

Lemma 1 (Good and Toulmin, 1956) *Let X be distributed as*

$$P(X = i) = \pi_i, \quad i \in \{0, 1, 2, \ldots, m\}, \tag{8}$$

where $\pi_i \geq 0$, $i \in \{0, 1, 2, \ldots, m\}$, and $\sum_{i=0}^{m} \pi_i = 1$. Then

$$\pi_i = \sum_{j=i}^{m} \frac{(-1)^{j-i}}{i!(j-i)!} E(X^{(j)}), \quad i \in \{0, 1, 2, \ldots, m\}. \tag{9}$$

- 周辺の確率関数を周辺階乗モメントで定める。

モメント逆転による LCCP 分布の周辺分布

命題 5 ($S_1, \ldots, S_n$ が母数 $(\mu, \boldsymbol{q})$ の *LCCP* 分布に従うとする。このとき

$$\Pr(S_i = s) = \sum_{j=s}^{\lfloor n/i \rfloor} \frac{(-1)^{j-s}}{s!(j-s)!} n^{(ij)} {q_i}^j \mu^j \frac{B_{n-ij}(\mu \boldsymbol{x})}{B_n(\mu \boldsymbol{x})}, \tag{10}$$

$$s \in \{0, 1, 2, \ldots, \lfloor n/i \rfloor\}.$$

例）Ewens 分布

- Ewens 分布の階乗モメントは Sibuya (1993) が与えている。結果として

$$\Pr(S_i = s) = \sum_{j=s}^{\lfloor n/i \rfloor} \frac{(-1)^{j-s}}{s!(j-s)!} \frac{\kappa^j \kappa^{[n-ij]} n^{(ij)}}{\kappa^{[n]}} (\frac{1}{i!})^j,$$

$$s \in \{0, 1, 2, \ldots, \lfloor n/i \rfloor\}.$$

Final Remark

- 社会的要請を割合うまく翻訳、定式化していると思われる例を紹介した。
  - 数学的にハードな問題は解いていない。
  - 格言「定式化した時点で問題は 7 割解けている」— 問題を 7 割理解した時点で定式化できた（後知恵）。
- 翻訳結果の技術は意味を説明でき、人を説得しやすい。
- 技術を社会的に利用するには人を説得しなければならない。