データマイニングと個人情報保護
Privacy Preserving Data Mining

菊池 浩明 †
Hiroaki Kikuchi

1. まえがき

情報の徹底的な電子化とインターネットの普及に伴い，個人の嗜好や購買記録などの大規模なデータベースの構築が進行している．特に，年収や家族構成などの顧客情報は，市場規模と宣伝効果を評価し効果的な商品販売計画を立てる観点から有用視されている．加えて，大規模なデータから特徴を抽出するデータマイニング技術と，CRM(Customer Relationship Management) と呼ばれる顧客情報管理技術の発達がその傾向を加速している．

ところがその一方で，個人情報の漏洩事件は後を絶たない．2004 年になってからでもメーカー，銀行，通信販売ビジネス，インターネットサービスプロバイダから，それぞれ 7 万 5 千，12 万，30 万，66 万，451 万件の個人情報が流出した．いずれも，住所氏名などの基本情報に電話番号や契約内容などの情報が漏れたとされている．これらの情報漏洩に対して，入退出管理や個人情報管理のポリシーを定めることで防止しようとするコンプライアンスルール（法令遵守）の導入も進んでいる．しかしながら，前述の流出事件の多くは，秘密を管理するオペレータなどの内部犯による犯行であり，その防止にはルールだけでは限界があることも認識されている．

結局の所，個人情報データベースの解析から得られる知識の有用性とそこから生じる流出事件によるリスクという危険性が互いに矛盾しているところにこの問題の本質がある．そこで，情報セキュリティ技術における秘密関数計算プロトコルを適用し，個人の属性情報を暗号化したまま解析することで，各種の統計情報や属性間の相関関係などの有益な知識を獲得しようとする研究が始まっている．Privacy-Preserving Data Mining (プライバシーを保護したデータマイニング）と呼ばれる一連の研究である．データマイニングの得られるものは母集団全体に成立する緩やかな相関ルールや論理決定木であって，個々の個人データが必要なわけではない．従って，個別のデータを秘匿したままデータマイニングが出来ればこの矛盾を解消することが出来そうである．この技術によって，次のような応用が考えられる．

- 顧客の購買情報を有する複数の商店が，互いの購買情報を秘匿したままデータマイニングをすることで，顧客全体の嗜好や購買に関する秘匿性をより高精度に抽出する．
- 容疑者情報を持つ警察と顧客の個人情報と運用ログを持つプロバイダーが互いの秘密（容疑者とログ）を秘密にしながら，その容疑者がそのプロバイダの顧客であるかどうかを同定する．
- 個人の健康診断データを秘匿したまま，特定の病気の発病に関する統計量を調査する疫学研究．
- DNA 情報を秘匿したまま，遺伝的な性質と DNA 塩基との関係を導出する．
- カンニングの常習犯を，学生のプライバシーを守ったままで同定して教育的な指導を行いたい複数の教員．

これらの応用を目的として，2004 年 3 月には米国 DIMACS/PORITA Workshop on Privacy-Preserving Data mining が開催され，約 25 件の研究発表が行われている [5]．これらの多くは公開鍵暗号技術を用いた暗号プロトコルを構成しているが，必ずしも暗号に限らないプライバシー技術もある．Verykios らは，用いられている技術の観点で，研究を次の 3 種類に分類した [4].

1. Heuristic-Based Techniques
   個人情報の変更や衛生化 (sanitization) によって，一般的なデータとして解析する試み．
2. Cryptography-Based Techniques
   秘密分散やセキュア関数計算 (Secure Multiparty Computation) によって，個人情報を秘匿したまま計算する試み．
3. Reconstruction-Based Techniques
   データに意図的なランダムノイズを乗せて，個人情報を意味のないものにゆがませる．ベイズの定理などに基づいて真のデータの分布を復元 (reconstruction) する試み．

本稿では，代表的なプライバシー保護データマイニングの手法として，2000 年に “Privacy-Preserving Data Mining” という同じタイトルで発表された全く異なる二つの研究 [3] と [2] を紹介する．興味深いことに，数あるデータマイニングアルゴリズムの中でも，これらは二つとも決定木学習アルゴリズムを取り上げていて，[3] は 2 番目，[2] は 3 番目に分類される．後者の著者 Agrawal はデータマイニングという分野を確立したアプリオリアルゴリズムの提案者でもある．

2. 準備

2.1 決定木学習アルゴリズム

学習データを $W = \{w_1, \ldots, w_m\}$, n 個のキーワードから成る部分集合を $K = \{k_1, \ldots, k_n\}$ とする．あるページ w_i にキーワード k_j が含まれるとき $a_j = 1$，含まれないとき 0 と置いて定義される $\boldsymbol{a}_i = (a_1, \ldots, a_n)$ を特徴ベクトルと呼ぶ．識別者は特徴ベクトルだけからページの識別を行うと仮定し，ある識別群に識別されるとき $f(\boldsymbol{a}_i) = 1$ と表す．特徴ベクトルの組 $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$ と，写像 $f : A \to \{0, 1\}$ を学習データと呼ぶ．

†東海大学電子情報学部

識別問題は，学習データ A が与えられたとき，A に対して最も誤差を小さくする論理式 f を見つける問題である．

f の**エントロピー** は $H(f) = -p_1 \log p_1 - p_0 \log p_0$ で与えられる，ただし，p_1 と p_0 は,$p_1 = |f^{-1}(1)|/m$, $p_0 = 1 - p_1$ で与えられる 1 と 0 の生起確率である．

キーワード k_1 が与えられたとき，k を含む集合 $A|_{k_1=1} = \{\boldsymbol{a} \in A | a_1 = 1\}$ について，$f|_{k_1=1}(a_2, \ldots, a_n) = f(1, a_2, \ldots, a_n)$ で定義される $n-1$ 変数の関数を，$f|_{k_1=1}$ とおく．同様に，$A - A|_{k_1=1}$ について定義される関数を，$f|_{k_1=0}$ とする．キーワード k に対する**情報利得** とは、

$$I(f;k) = H(f) - E[H(f|k)]$$

で定義される値である．キーワード k による識別で期待されるエントロピーの削減量を表している。ただし、

$$E[H(f|k)] = \sum_{x=0,1} P(k=x) H(f|_{k=x})$$

である．ID3[7, 6] では, 全ての変数について利得を計算し, 最も大きな利得が得られるキーワードについて f を展開する．その結果生じる 2 つの $n-1$ 変数の関数 $f|_{k=0}$ と $f|_{k=1}$ の各々に、同じ手続きを再帰的に適用していき, 定数 0 または 1 になるまで繰り返す．

2.2 データマイニング

現在よく用いられているデータマイニング技法には，アプリオリアルゴリズムによる相関ルール抽出，同数値属性を対象とした相関ルール抽出，エントロピー利得に基づく決定木アルゴリズム，同数値属性を対象とするアルゴリズム，ユークリッド距離を用いたクラスタリング手法などがある．

2.3 秘密関数計算

基本とする秘密関数計算の技術に関しては，古くは 80 年代後半には既に任意の関数が入力値を伏せたまま計算できることが証明されていた．しかし，そのためには膨大な計算量と通信量が必要であり机上の空論であった．それが，暗号プロトコルの最近の研究成果により，総和や最大値などの特定の関数に限っては，より現実的なコストで実現できることが知られてきて，例えば，電子選挙や電子入札などへの応用が始まってきた．データマイニングなどの分野で利用されている多属性の決定木解析などのより複雑な解析手法に秘密関数計算を適用する．

秘密関数計算には，(1) 秘密分散法に基づくマルチパーティプロトコル，(2) 準同型性を満たした公開鍵暗号による単一サーバによる暗号プロトコル，(3) 匿名通信路や紛失通信路プロトコル (Oblivious transfer) を応用したものなどがある．

3. プライバシ保護データマイニング

3.1 モデル

本モデルは次のとおりである．

1. 個人ユーザ U_i は信頼できる鍵管理者 A の公開鍵 PK で属性情報 a_i を暗号化して $E_A(a_i)$ データベースに登録する．ここで，$E_A(x)$ はデータ x の A の公開鍵による暗号化，$D_A(x)$ は同復号化を表す．

2. データ解析者 C は，復号化の為の秘密鍵を持たないが，n 個のデータ $E_A(a_1), \ldots, E_A(a_n)$ を暗号化されたまま解析を行ない，解析結果 $E_A(f(a_1, \ldots, a_n))$ を出力する．(その過程で鍵管理者 A と通信を行う可能性がある)．

3. 解析結果は暗号化されており，個人情報とは切り離されて鍵管理者 A に渡され，データマイニングの結果 $y = D_A(E_A(f(a_1, \ldots, a_n)) = f(a_1, \ldots, a_n)$ (例えば，相関ルールや決定木) だけが復号化される．

従って，営利活動からの市場調査の必要性と個人情報の保護という相反する二つの要請がこのモデルによって満たされる．そして，安全なデータマイニングが実用的なコストで実現できることが示されれば，歯止めの利かない個人情報の漏洩に対して大きな社会的な貢献となることが期待できる．

3.2 Cryptographic-Based Approach [3]

Lindel と Pinkas によって提案されたアルゴリズムである．A と B の二つの組織が，それぞれの学習データ f_A と f_B を秘密にしたままで，n 個ある属性（キーワード）の中から最も情報量利得を最大化する属性 k^* を見つけることが問題である．

例えば，表 1 のビールを購入した顧客についての性別, 年齢，車所有の 3 種類の属性情報からなる学習データを考えよう．商店 A と B がそれぞれ，3 件づつの学習データ f_A, f_B を有している時，これらを合わせて決定木でビールをよく買いに来る客の特徴をつかみたい．

表 1: 学習データ

No	性別	年齢	車所有	ビール購入
1	F	20	Y	0
2	M	10	N	1
3	M	30	Y	1
f_A				
4	F	30	Y	1
5	F	10	N	0
6	M	20	N	1
f_B				

決定木学習の目的は，ターゲットとなる属性（クラス）のエントロピーを最も下げる属性を見つけることである．もしも，車の所有で 6 件のデータを分類していくと図 1 の A の決定木が得られる．一方，先に性別で分類すると同図の B の木が得られる．明らかに，A よりも B の木の方が深さが浅く，節点数も少ないシンプルな木である．別の言い方をすれば，属性「性別」は「車所有」よりも，エントロピーを下げる「よい識別の属性」である．

従って，二つの商店がそれぞれの顧客情報を秘匿したまま，

「性別」>「車所有」

をテストできればよいことになる．この順序関係は情報量利得で計ることが出来，例えば，車を所有しているという条件付エントロピーは，A の 1 と 3 の二人と B の

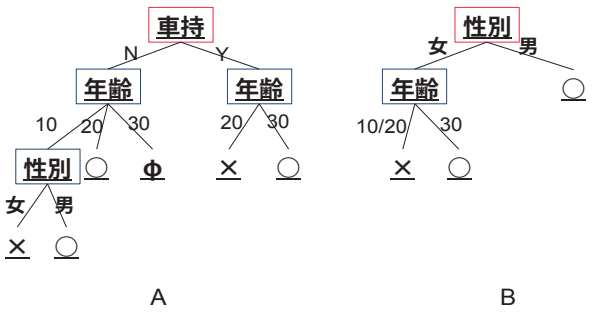

図 1: 「ビール購入者」を表した決定木の例

4 の計 3 人分の情報から,

$$H(\text{ビール} \mid \text{車} = Y) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log\frac{3}{1}$$

で与えられる．しかしながら，ここで A に車という条件で 2 人，B に 1 人という情報は互いに漏らしてはならない．結局のところ，この問題は，$x_A + x_B = x$ となる x_A, x_B を秘密にした A と B が，$y_A + y_B = x\log x$ となる y_A と y_B を求める問題に帰着する．

これを計算する為には，いくつかの暗号プロトコルと紛失多項式評価 (OPE: Oblivious Polynomial Evaluation) と呼ばれるテクニックを用いる．OPE は，A が秘密の多項式 $P()$ を定義し，B が入力 x を秘密にしたままで $P(x)$ の値を計算するプロトコルである．これを用いると，$l_A + l_B = \log x, x_A + x_B = x$ となる x_A, l_A を持った A と x_B, l_B を持った B とが，次のようにして $y_A + y_B = x\log x$ を計算することが出来る．

まず,

$$\begin{aligned} x\log x &= (x_A + x_B)(l_A + l_B) \\ &= x_A l_A + x_A l_B + x_B l_A + x_B l_B \end{aligned}$$

なので，A は $x_A l_A$ を，B は $x_B l_B$ をそれぞれ独立に計算する．残りの項を求める為に，A は $P_1(z) = x_A z + r_1$, $P_2(z) = l_A z + r_2$ という二つの一次式を作る．ここで，r_1, r_2 は乱数である．次に，B は OPE を用いて，l_B と x_B を秘密にしたまま，$P_1(l_B)$ と $P_2(x_B)$ を A に計算してもらう．こうして計算された $y_A = x_A l_A - r_1 - r_2$ と $y_B = x_B l_B + P_1(l_B) + P_2(x_B)$ は次の様に，正しい分散値になっている．

$$\begin{aligned} & y_A + y_B \\ &= x_A l_A - r_1 - r_2 + x_A l_B + l_A x_B + r_1 + r_2 + x_B l_B \\ &= x\log x \end{aligned}$$

こうして，A, B は協力して，上のプロトコルを繰り返しながら全ての属性についての情報量利得を求め，最適な識別を行う属性を決定していく．

以上の計算の過程において，f_A, f_B に関する情報は 1 ビットも漏らさないことが，ここでいうプライバシ保護である．ただし，得られた決定木から互いの持つ秘密についてはある程度分かってしまうことにも注意が必要である．例えば，A の年齢に関するエントロピーが最大（つまり，ビールの購入には年齢は独立で，識別には全く貢献していない）であったのに，その属性が木のルートに選ばれていたとすると，A はそのことから B の分布に年齢が大きく貢献していたことを知ってしまう．

3.3 Reconstruction-Based Approach [2]

Agrawal and R. Srikant によって発表されたアルゴリズムである．暗号プロトコルによるアプローチが情報量を一切もらさずに，正確に知識を計算しようとするのに対して，この試みでは，誤差とプライバシーの守り方に度合いを与え，統計的に母集団での規則性を導こうとしている．

基本原理は，個人情報に意図的にランダムノイズを乗せて，プライバシーを保護しようとするものである．真の値を $x_1, x_2, \ldots, x_n$ とする．ここに加える確率分布 Y の乱数を $y_1, y_2, \ldots, y_n$ とする時，再現問題 (Reconstruction Problem) とは，$w_1 = x_1 + y_1, w_2 = x_2 + y_2, \ldots, w_n = x_n + y_n$ の値と確率変数 Y から，真の値 X の確率分布を見積もることである．例えば，年齢が $x = 20$ 代であるという個人情報をそのまま渡す代わりに，一様分布（またはガウス分布）の乱数 r を加え，$x + r = 30$ のように歪んだ値を登録する．30 という属性値を持った顧客がいても，本当に 30 代なのか乱数で 40 代から歪まされたのか，第三者には区別がつかない．

提案アルゴリズムでは，ここに，ベイズの定理を用いて，次の式で逐次的に確率密度関数 f_X を推定する．

$$f_X^{j+1}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X^i(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X^j(z) dz}$$

ここで，初期値 f_X^0 は一様分布とする．確率密度がわかれば，属性ごとの情報量利得を求めることができるので，目的とする決定木の学習される．

Y の大きさ（ランダマイズの強さ）によって決まる同定精度とプライバシーの間にはトレードオフがあり，Y の分散を大きく取るとそれだけ誤差が広がるがプライバシーに関する安全性は高くなる．

4. その他のプライバシー技術

4.1 リング署名

リング署名は，2001 年の ASIACRYPT で Rivest, Shamir, Tauman らによって提案された暗号プロトコルである [1]．“How to leak a secret” (秘密を漏らす方法) という挑発的な論文名をつけている．このプロトコルは次の二つを目的としている．

- メッセージの完全性 (メッセージが署名されてから改竄されていないことの確認),
- 署名者の匿名性 (署名者が列挙されたものの中の誰かであるかわからない)

リング署名は，署名者が他のユーザの公開鍵と自分の秘密鍵を用いて連鎖的に計算するディジタル署名である．乱数から初めて，他の候補者の公開鍵を用いて署名の連鎖を作り，最後に自分の秘密鍵を用いて，署名連鎖と最初の乱数の逆関数を求めて，リングを「つなぐ」．この

署名の連鎖と署名候補者の公開鍵の列がリング署名となる．検証者は，リング署名から，少なくともその候補者の中の一人がリングを「つない」でいることを確認するが，誰が真の署名者かはわからない．署名の連鎖がどこから始まっているかわからないことが匿名性の保証である．戦国時代に，誰が発起人かわからないように傘状に連判を行った「唐傘連判」こそがリング署名の原理であるとも言える．

以下に，公開鍵（トラップドア付き一方向性関数）を用いて抽象化した，3 ユーザ (A, B, C) のリング署名の具体例を次に示す．検証者 V は任意の第三者である．

1. 署名者 B は，乱数 R を選び，$c_C = R$ とする．
2. B は乱数 s_C を選び，候補者 C の公開鍵 E_C を用いて，$c_A = E_C(R, s_C)$ を求める．
3. B は同様に，$c_B = E_A(c_A, s_A)$ を求める．
4. B は，最後に，自分の秘密鍵 E_B^{-1} を用いて，$s_B = E_B^{-1}(c_B, R)$ を求めて，リングを閉じる．リング署名は，$s_A, s_B, s_C, c_A, c_B, c_C$ である．
5. 検証者 V は，リング署名から，$c_B \stackrel{?}{=} E_A(c_A, s_A)$ を満たすことを順に検証する．

2001 年の Rivest らの発表を元に，次のような拡張や一般化の改良が行われている．

- ハッシュ関数と一方向性関数を用いた一般化と効率化（阿部，大久保ら，2002 ASIACRYPT)
- しきい値リング署名（n 人中の k 人以上が協力して行う署名) (Bresson, Stern and Szydlo, 2002 Crypto, 桑門，田中, 2002 ISEC, 菊池，多田，2002 CSS)
- 匿名性破棄 (Camenish and Lysyanskaya, 2002, 中西, 2003 SCIS)

リング署名に限らず，暗号プロトコル一般に広く言える課題はその効率である．プライバシーを保証するためには，通常のプロトコルに対して必ず何らかの対価を納める必要がある．理論的な興味だけで構築されたプロトコルの中には，これらのコストが膨大で非現実的なものも少なくない．

- ラウンドコスト　（プレーヤー間で行わなくてはならない通信の回数）
- 通信コスト　（一回の通信で消費する帯域の大きさ）
- 計算コスト　（各プレーヤーが実行しなくてはならない計算量および暗号化にかかる処理）

これらのコストは，署名候補者の数 n の関数で与えられる．基本リング署名のコストは，通信と計算のコストとも $O(n)$，ラウンド数は 1 である．

5. おわりに

個人情報流失事件での犯行手口は，外部からの不正侵入ではなく，契約社員などの内部犯による正規のユーザによる裏切りによるものがほとんどであるという．ひどいものになると，アクセス履歴（ログ）が数週間で処分されてしまっていて，犯人がいつことを成したのかさえ同定できないと言うお粗末なものさえあったという．いかに数学的に美しく高度な認証プロトコルを発明しても，高速で精度の良い不正侵入検出システムを開発していても，これではまるで役に立っていない．最先端の技術を誰にでも使える一般的な技術に成熟させ，末端にまで浸透するまで我々研究者は責任を持たなくてはなるまい．

参考文献

[1] R. L. Rivest, A. Shamir and Y. Tauman, “How to Leak A Secret”, *ASIACRYPT 2001*, LNCS 2248, Springer, pp.552-565, 2001.

[2] R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,” Proc. of the ACM SIGMOD Conference on Managament of Data, pp. 439-450, 2000.

[3] Y. Lindel and B. Pinkas, “Privacy Preserving Data Mining,” Journal of Cryptology, Vol. 15, No.3, pp. 177-206, 2002.

[4] V. S. Verykios, E. Bertino and I. N. Fovino, “State-of-the-art in Privacy Preserving Data Minig,” SIGMOD, Vol. 33, No. 1, 2004.

[5] DIMACS/PORITA Workshop on Privacy-Preserving Data Mining (http://dimacs.rutgers.edu/Workshops/Privacy/program.html)

[6] Tom Mitchell, Decision Tree Learning, *Machine Learning*, McGraw-Hill, pp.52-79, 1997

[7] Quinlan,J.R., Induction of decision trees, Machine Learning, 1(1), pp.81-106, 1986

[8] 沼尾雅之,「セキュリティと AI」，人工知能学会誌, Vol. 19, No. 2, 2004.