匿名ゲノムデータベースに対する連鎖不均衡を用いた脱匿名化攻撃の提案と評価

高山 晃一[1,a)]　荒井 ひろみ[2]　中川 裕志[2]

概要：個人のゲノムデータは，その所有者のプライバシを大きく侵害しうる情報を多く有する極めて秘匿性の高いデータである．一般には，プライバシ保護のため，ゲノムデータは個人の識別子を削除した匿名化状態で取り扱われることが多い．しかし，このような単純な匿名化は，攻撃対象となる個人に対する背景知識を持った攻撃者に対する安全性を保証できない．我々は，個人のゲノムデータの一部を知った攻撃者が匿名ゲノムデータベースに対する脱匿名化を行う攻撃に着目する．その攻撃モデルの下で，最尤推定と二値分類を用いた攻撃モデルを提案し，実データを用いた実験によりその攻撃精度を評価する．

キーワード：PWS，プライバシ，脱匿名化，ゲノム

1. はじめに

個人のゲノムデータは，個人の疾病リスクや体質などのセンシティブな情報を多く含んでいる極めてプライベートな情報である．こうしたセンシティブな情報は医療の質の改善などに有益であると期待されており，ゲノムデータの分析ニーズは高まってきている．しかし，ゲノムデータにはネガティブな一面も存在する．例えば，個人のゲノムデータが何らかの形で特定されると，人種差別や就職差別，ないしは保険加入の制限などがその個人に対して行われる危険性がある．こうしたプライバシ侵害の危険性を受けて，ゲノムデータについてのプライバシリスクの評価，ならびにプライバシ保護の重要性が広く認識され，近年盛んに研究が行われている．

一般に，ゲノムデータは個人が特定されないよう個人の識別子を削除した匿名化状態で保管されていることが多い．本研究ではこうしたデータベースを匿名ゲノムデータベースと呼ぶ．

本研究では，匿名ゲノムデータベースに対する脱匿名化攻撃を扱う．特に，データベースに含まれるゲノムデータが個人の SNP である場合を考える．なお，SNP とは個体間で個人差が生じうるゲノムデータのことを指す．また，脱匿名化攻撃としては，攻撃対象となる個人の SNP を，その個人の SNP が含まれる匿名ゲノムデータベースの中から特定するという問題設定に限定して議論を行う．悪意のある攻撃者がこうした脱匿名化攻撃に成功すると，前述したような様々なプライバシ侵害が生じうる．したがって，脱匿名化攻撃のリスクを適切に評価することが必要となる．

脱匿名化攻撃に注目した研究としては，以下のものが存在する．Lin らは，攻撃対象の個人に関する知識として，匿名ゲノムデータベースに含まれる約 100 個の SNP を知った攻撃者は，単純なマッチング攻撃を用いて高い確率で個人を一意に特定できることを示した [3]．また，Humbert らは，攻撃対象の個人の目の色などの形質データを利用した脱匿名化攻撃を提案した [1]．彼らは，攻撃対象の個人の形質データと，ゲノムデータと形質データの間の相関関係についての情報を攻撃者の背景知識として設定し，最尤推定に基づいた予測手法を提案した．この研究により，攻撃者は，攻撃対象の個人に関して匿名ゲノムデータベースに含まれている SNP の情報を全く知らない場合であっても脱匿名化攻撃を成功しうることが示唆された．

これらの既存研究では，様々な問題設定の下での攻撃モデルの評価が行われているが，まだ想定されていない攻撃のシナリオも存在する．本研究では，攻撃者の背景知識に着目し，これまで想定されていなかった問題設定の下での攻撃モデルを提案する．まず攻撃者の背景知識を以下の二種類に分類する：(a) 攻撃対象の個人についての知識 (b) ゲノムデータに関する補助知識．また，攻撃者が攻撃対象の個人について知っていて，かつ匿名ゲノムデータベースにも含まれている SNP の情報を “explicit knowledge” と定義した．一方で，攻撃者が攻撃対象の個人について知っ

[1] 東京大学大学院学際情報学府
[2] 東京大学情報基盤センター
[a)] k.takayama0902@gmail.com

	攻撃者の背景知識			手法
	(a): 攻撃対象の個人についての知識		(b): 補助知識	
	Explicit knowledge	Implicit knowledge		
Humbert et al. [1]	×	形質データ	ゲノム-形質間の相関	最尤推定
Lin et al. [3]	○	×	×	マッチング
本研究	○ - - - - - - - - - - ×	SNP	・SNP の頻度 ・SNP 間の相関	・最尤推定 ・二値分類

表 1: 匿名ゲノムデータベースに対する脱匿名化攻撃を対象とした既存研究の比較．各行はそれぞれの研究に対応する．なお，本研究では explicit knowledge を持つ攻撃者と持たない攻撃者の両方の場合を同じ枠組みで扱うことが可能であるため，破線でその旨を示した．

ていて，かつ “explicit knowledge” に含まれていない情報を “implicit knowledge” と定義した．これらの概念を用いて既存研究を整理すると，表 1 に示したように，Lin らの研究 [3] は，(a) の知識が explicit knowledge のみであり，(b) の知識は何も持たない場合の攻撃に相当する．また，Humbert らの研究 [1] は，(a) の知識が攻撃対象の個人についての形質データ（これは implicit knowledge である）であり，(b) の知識がゲノムと形質間の相関関係についての情報である場合の攻撃に相当する．我々の知る限り，explicit knowledge を持たない脱匿名化攻撃は，Humbert らの研究 [1] を除いて存在しない．そこで，本研究では，(a) の知識が攻撃者の個人についての SNP の一部であり，(b) の知識が SNP の頻度情報と SNP 間の相関情報である場合を想定した攻撃モデルを提案する．

また，本研究では，こうした攻撃者の背景知識に関する問題設定に合わせた攻撃手法の提案も行う．まず，Humbert らの研究 [1] で使用されていた最尤推定による攻撃手法を基にしつつ，SNP の頻度情報と SNP 間の相関情報を利用するよう改良を施した手法を提案した．次に，脱匿名化攻撃を二値分類問題に帰着させ，決定木による分類を行う手法を提案した．さらに，二値分類を行う前処理として次元削減を行う手法を提案した．我々の知る限り，匿名ゲノムデータベースに対する脱匿名化攻撃を二値分類によって行った研究は本研究の他には存在しない．

提案手法の評価のため，実データを用いた実験を行った．結果から，提案手法は様々な設定の攻撃者に対してランダムな予測より十分に高い精度で脱匿名化に成功することが分かった．特に，二値分類を用いた攻撃手法が顕著な精度向上を達成した．

2. 準備

本章では，本研究に用いたゲノムデータについて説明する．

ヒトの DNA は {A,T,C,G} の四種類の塩基からなる塩基配列として表現され，そのうち A と T, C と G がそれぞれペアになって二重螺旋状に並んでいる．DNA は，染色体と呼ばれる高分子の中に含まれており，各染色体は DNA を一本ずつ中に含んでいる．ヒトの場合，染色体が 23 対（46 本）存在しており，各対における染色体は両親から一本ずつ受け継がれる．さらに，単一の染色体が持つ塩基配列のことを haploid genotype と呼び，一対の染色体の塩基配列の組を diploid genotype と呼ぶ．塩基配列上で各塩基が存在する場所のことを座位と呼ぶ．

ヒトには約 30 億個の塩基が含まれている．そのうち，ほとんどの塩基は全人類で共通しているが，約 0.3% の塩基には個人差があると言われている．特に，単一の塩基が一定以上の頻度で変異を起こしている場合，その塩基を一塩基多型（Single Nucleotide Polymorphism; SNP）と呼ぶ．なお，ほとんどの SNP は二種類の塩基しかとりえないと言われており，他の多くの研究と同様，本研究でもそうした SNP のみを扱う．各 SNP のとりうる二種類の塩基のうち，高頻度で現れるものを major allele, 低頻度で現れるものを minor allele と呼ぶ．また，本研究では，各染色体における SNP が diploid genotype として表現される場合を考える．このとき，major allele を B，minor allele を b と表記すると，各 SNP は $\{BB, Bb, bb\}$ の三つのうちどれかの値をとることになる．以降，BB を 0, Bb を 1, bb を 2 と表記する．

複数の SNP 間の関係性に着目すると，各 SNP のとる値には相関が見られる．こうした現象は連鎖不平衡と呼ばれている．特に，二つの SNP 間の相関の強さに相当する統計量（r^2 と表記する）がしばしば考慮される．また，このようなペアワイズ相関に加え，近年は高次の相関情報も考慮されている [2].

3. 攻撃モデル

本研究では図 1 のような攻撃モデルを考える．本章では，攻撃者の知識と脱匿名化攻撃についての定式化を行った後，攻撃シナリオについての考察を行う．

3.1 攻撃者の知識

本節では，攻撃者が脱匿名化攻撃を行うために使用する知識について整理する．

まず前提として，本研究では，攻撃者は攻撃対象となる匿名ゲノムデータベース D_X を直接観測できると仮定する．D_X は n 人の個人の SNP についての情報を含んでお

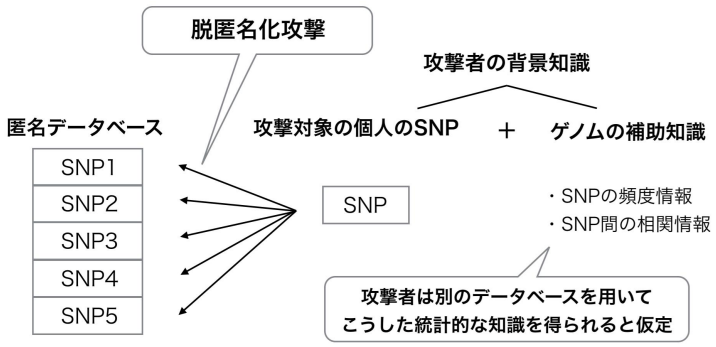

図 1: 攻撃モデルの概要．攻撃者は，匿名ゲノムデータベースに含まれる全てのレコードにアクセスし，どのレコードが攻撃対象の個人によって生成されたものであるかを推定する．攻撃者は，攻撃対象の個人のついての知識（本研究でのは SNP）と，ゲノムデータに関する補助知識（本研究では SNP の頻度と SNP 間の相関）という二種類の背景知識を持つと仮定する．

り，$D_X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ と表記されるとする．また，D_X に含まれる SNP の座位の集合を $J_X \subset \mathbb{N}$ と表記する．このとき，D_X に含まれる個人 $i \in \{1, \cdots, n\}$ の SNP は，$\boldsymbol{x_i} = \{x_{i,j}\}_{j \in J_X}$, $x_{i,j} \in \{0, 1, 2\}$ と表記できる．

次に，D_X 以外に攻撃者が持つ知識を攻撃者の背景知識と呼ぶこととし，以下のようにその詳細を整理する．1 章で述べたように，我々は攻撃者の背景知識を以下の二種類に分類した: (a) 攻撃対象の個人についての知識 (b) ゲノムデータに関する補助知識．図 1 と表 1 に示したように，本研究では，(a) の知識が攻撃者の個人についての SNP の一部であり，(b) の知識が SNP の頻度情報と SNP 間の相関情報である場合を対象としている．

このとき，攻撃者の強さ，すなわち攻撃成功確率の大きさは，攻撃者が持つ攻撃対象の個人についての知識に依存すると考えられる．ここで，攻撃対象の個人について攻撃者が知っている SNP を $\boldsymbol{y}$ と表記し，$\boldsymbol{y}$ に含まれる SNP の座位の集合を $J_Y \subset \mathbb{N}$ と表記する．すると，$\boldsymbol{y} = \{y_j\}_{j \in J_Y}$, $y_j \in \{0, 1, 2\}$ と表記できる．このとき，例えば, 攻撃者の強さは J_X と J_Y の積集合の要素数 $|J_X \cap J_Y|$ が多いほど強くなることは明らかである．なお，J_X の補集合を $\bar{J_X}$ としたとき, $J_X \cap J_Y$ は explicit knowledge であり，$\bar{J_X} \cap J_Y$ は implicit knowledge である．本研究では簡単のため，$|J_X \cap J_Y| = 0$ の場合のみを考えるが，$|J_X \cap J_Y| > 0$ の場合への拡張も可能である．また，$|J_X|$ や $|J_Y|$ が大きいほど，攻撃者は強くなると考えられる．

さらに，攻撃者の強さは，攻撃者が持つ SNP 間の相関についての補助知識にも依存すると考えられる．本研究では，攻撃者は，ペアワイズ相関についての統計量 r^2 に加え，D_X とは異なるゲノムデータベース D_O にアクセスでき，それらから補助知識を抽出することを仮定する．D_O としては，$D_O = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_{tr}}\}$, $\boldsymbol{z}_i = \{z_{i,j}\}_{j \in J_X \cup J_Y}$ $(i \in \{1, \cdots, N_{tr}\})$ として表されるものを仮定する．このとき，例えば， 相関の情報 $r^2(j_x, j_y)$, $j_x \in J_X, j_y \in J_Y$ は攻撃者の強さを示す指標として有用であると考えられる．このような J_X と J_Y の間の相関関係は図 2 のような二部グラフを用いて表すことができる．なお，ここでは J_X 内，J_Y 内の相関関係は考えない．この二部グラフでは，各ノードは各 SNP を表し，各エッジは二つの SNP 間の相関を表している．また，r^2 統計量が公開されていないペアは相関が弱いと考え，そうしたペアについてはエッジを消去する．一般に，強い相関を持つペア (j_x, j_y) の数が多いほど，攻撃者は強くなると考えられる．

本研究では，こうした背景知識に関する問題設定の下で，様々な強さをもった攻撃者を想定して脱匿名化リスクを評価することを試みる．特に，J_X と J_Y の相関関係を利用した脱匿名化攻撃を評価対象としているため，J_X と J_Y の相関関係が多様になるような設定で評価実験を行う．実験設定の詳細は 5 章で説明する．

3.2 脱匿名化攻撃

本節では脱匿名化攻撃の定式化を行う．攻撃者は，3.1 節で述べた背景知識を基にスコア関数 $f(\cdot, \cdot)$ を構築し，その関数を用いて攻撃を行うとする．f は，二つのベクトル $(\boldsymbol{x}_i, \boldsymbol{y})$ を入力とし，実数値を出力するような関数とする．高い攻撃精度を達成するには，f は以下のような性質を満たすことが望ましい: 二つの入力ベクトル $\boldsymbol{x}_i$ と $\boldsymbol{y}$ が同一人物の SNP であった場合は高い値を出力し，そうでなかった場合は低い値を出力する．なお，関数 f の具体的な構築方法については 4 章で説明する．

攻撃者は，f を構築した後，以下のように攻撃を行う: D_X 内の各個人 $i \in \{1, \cdots, n\}$ に対して $f(\boldsymbol{x}_i, \boldsymbol{y})$ を計算し，その値が最大となる SNP を持つレコード $\hat{i} = \arg\max_{i \in \{1, \cdots, n\}} f(\boldsymbol{x}_i, \boldsymbol{y})$ を攻撃対象の個人のレコードとして予測する．

3.3 攻撃シナリオ

本章で説明してきた攻撃モデルは，[1] と同様に，様々な攻撃シナリオに適応することが可能である．例えば，保険料金の値上げやサービスの制限を目的にした保険会社が，攻撃対象の個人の SNP の特定を試みる場合である．そう

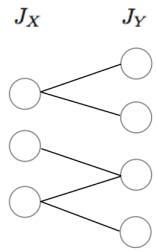

図 2: J_X と J_Y の間の相関関係を表現した（無向）二部グラフの例.

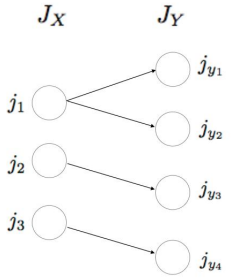

図 3: 二部グラフの別の例. 有向グラフを仮定し，J_Y 内の各ノードは高々一つのエッジしか持たないモデル.

した保険会社が出現すると，個人の実名つき SNP の一部や匿名ゲノムデータベースを売買するブローカーが出現する可能性も高まり，プライバシ侵害リスクがさらに高まる恐れがある．また，患者の SNP 情報の一部を知った医療関係者が，自身のアクセスできる匿名ゲノムデータベースを利用して “好奇心から” 脱匿名化を試みる場合も考えられる．こうした攻撃は，自身の情報を不当に知られたくないという意味での患者のプライバシを侵害している．

なお，ここまでの説明では，既存研究 [1] と同様に，攻撃者は攻撃対象の個人のデータが匿名ゲノムデータベースに含まれていることを知っていることを暗に仮定してきた．こうした仮定が現実的でない場合もあるが，提案手法で用いる関数の出力に適当な閾値を設定することによってその仮定を取り除くことが可能である．本研究では，より強力な攻撃者について考えるため，この仮定の下での攻撃についてのみ考える．

4. 手法

本章では，攻撃者が関数 f を構築する手法を提案する．まず最尤推定を用いる手法を提案し，次に二値分類を用いる手法を提案する．

4.1 最尤推定

本節では，関数 $f(\boldsymbol{x}_i, \boldsymbol{y})$ に尤度 $\Pr[\boldsymbol{y}|\boldsymbol{x}_i]$ を使用する手法を提案する．この手法は，攻撃対象の個人の形質情報を利用して脱匿名化攻撃を行う既存研究 [1] から着想を得ている．ここで，尤度の厳密な計算は困難であるため，[1] と同様に何らかの近似が必要である．本研究では，連鎖不平衡によって説明される SNP の出現パターンに基づいて尤度を近似する．すなわち，同時分布を SNP 間のペアワイズ相関に基づいて分解し，周辺分布と条件付き分布の積として表現する．

簡単のため，同時分布 $\Pr[\boldsymbol{y}, \boldsymbol{x}_i]$ が図 3 のような有向二部グラフで近似されることを仮定する．図 3 では，J_Y 内の各ノードは，J_X 内の一つのノードに依存している．この依存関係は SNP 間のペアワイズ相関についての r^2 統計量に基づき設定し，本研究では，各 $j_y \in J_Y$ について y_{j_y} はたかだか一つの SNP $x_{i,jmax(j_y)}$ に依存すると仮定した．座位 $jmax(j_y)$ は以下のように決定した:

$$jmax(j_y) := \arg\max_{j \in J_X} r^2(j, j_y)$$

すなわち，J_Y 内の各 SNP について最も相関が強い座位との相関のみを考慮した．図 3 の確率モデルでは，$jmax(j_{y_1}) = jmax(j_{y_2}) = j_1$ であり，(j_1, j_{y_1}, j_{y_2}) は部分グラフのうちの一つである．さらに，分解された部分グラフを要素とする集合を C とし，C の各要素 c について，$c_x := c \cap J_X$，$c_y := c \cap J_Y$ と定義する．このとき，最尤推定を用いる攻撃者は，攻撃対象の個人のレコードを以下のような $\hat{i}$ として予測する:

$$\begin{aligned}
\hat{i} &= \arg\max_{i \in \{1,2,\cdots,n\}} \Pr[\boldsymbol{y}|\boldsymbol{x}_i] \\
&= \arg\max_{i \in \{1,2,\cdots,n\}} \frac{\Pr[\boldsymbol{y}, \boldsymbol{x}_i]}{\Pr[\boldsymbol{x}_i]} \\
&\simeq \arg\max_{i \in \{1,2,\cdots,n\}} \frac{\prod_{c \in C} \prod_{j_y \in c_y, j \in c_x} \Pr[y_{j_y}|x_{i,j}] \Pr[x_{i,j}]}{\prod_{j \in J_X} \Pr[x_{i,j}]}
\end{aligned}$$

なお，c_x が空集合であれば，$\Pr[y_{j_y}|x_{i,j}] = \Pr[y_{j_y}]$; $^\forall j_y \in c_y$, $^\forall j \in c_x$ と計算する．最終行の近似は，確率モデルを図 3 のように仮定したことに由来している．

本研究では，攻撃者は D_X とは異なるデータベース D_O にアクセスできると仮定しており，攻撃者は D_O を用いて SNP 間の相関 $\Pr[y_{j_y}, x_{i,j}]$ と SNP の頻度 $\Pr[x_{i,j}]$ を計算する．なお，確率が 0 になることを防ぐために MAP 推定を行った．$\Pr[y_{j_y}, x_{i,j}]$ の事前分布にはディリクレ分布を，$\Pr[x_{i,j}]$ の事前分布にはベータ分布を使用した．事前分布のハイパーパラメータはすべて 2 と設定した．

4.2 二値分類

本節では，f に二値分類器 f_{bc} を使用する手法を提案する．攻撃者は，以下のような性質を持つ二値分類器の構築を試みるとする: $f_{bc}(\boldsymbol{x}_i, \boldsymbol{y})$ の出力は，二つの入力ベクトル $\boldsymbol{x}_i$ と $\boldsymbol{y}$ が同一人物の SNP であった場合は 1 に近い値を出力し，そうでなかった場合は 0 に近い値を出力する．すなわち，f_{bc} の目的は，二つの入力ベクトルか同一人物のものであるか否かを分類することとなる．この定式化により，SNP の頻度や，SNP 間の高次の相関情報を利用した攻撃が可能となる．また，この定式化を最尤推定に基づく手法と比較すると，二値分類はその目的が脱匿名化そのものになっているという点で自然な定式化であり，より高精度な脱匿名化リスクの評価を行えることが期待される．

4.2.1 特徴量の設計

脱匿名化リスクを高精度に行うためには，前処理におい

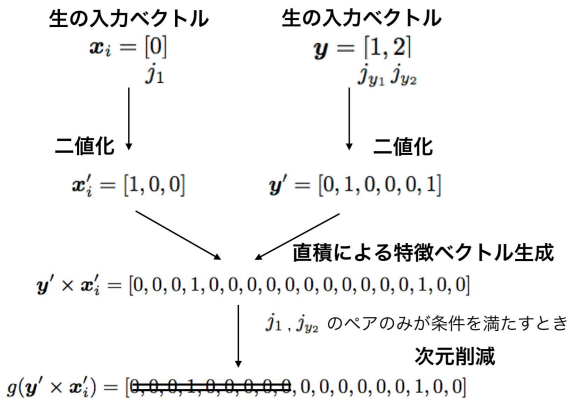

図 4: 特徴ベクトルの構築

て特徴量をどう設計するかが重要となる．図 4 と以下に示すように，本研究では特徴量の設計を三つの手順に分けて行う．

まず，二つの入力ベクトル $(\boldsymbol{x}_i, \boldsymbol{y})$ をそれぞれ二値特徴量に変換する: $\boldsymbol{x}_i \to \boldsymbol{x}'_i$, $\boldsymbol{y} \to \boldsymbol{y}'$．例えば，$\boldsymbol{y} = [1, 2]$ は $[0, 1, 0, 0, 0, 1]$ に変換される．変換後のベクトルの最初の三要素 $[0, 1, 0]$ は，変換前のベクトルの一つ目の要素 1 に相当し，変換後のベクトルの二つ目の三要素 $[0, 0, 1]$ は，変換前のベクトルの二つ目の要素 2 に相当する．

次に，変換後の二つの二値特徴量の直積 $\boldsymbol{y}' \times \boldsymbol{x}'_i$ をとることで 9 $|J_X|$ $|J_Y|$ 次元の特徴ベクトルを作成する．例えば，$\boldsymbol{x}_i = [0]$ のとき $\boldsymbol{x}'_i = [1, 0, 0]$ となり，$\boldsymbol{y} = [1, 2]$ に対して直積は図 4 の三行目のように計算される．この手続きの代わりに，$\boldsymbol{x}'_i$ と $\boldsymbol{y}'$ を結合することにより特徴ベクトルを作成すると，J_X と J_Y 間の組合せに関する相関情報が特徴量に反映されない恐れがある．今回の脱匿名化攻撃では J_X と J_Y 間の組合せの相関情報が重要であるため，直積をとることによりその情報を明示的に考慮した．

最後に，特徴ベクトルの次元削減を行う．$|J_X|$ や $|J_Y|$ が大きいとき，特徴ベクトルの次元は大きくなる．特徴ベクトルの次元が必要以上に大きい場合，過学習を引き起こす恐れがあるため，不要な次元を削減することにより性能が高まることが期待される．そこで，特徴ベクトルの各次元の基となった二つの SNP 座位のペアワイズ相関の統計量 r^2 を次元削減の尺度として用いた．本研究では，$r^2(j, j_y) \geq 0.1;\ j \in J_X, j_y \in J_Y$ を満たす座位に相当する次元のみを残した．図 4 の四行目では，(j_1, j_{y_2}) のペアのみがこの条件を満たしたときの次元削減を関数 g と定義し，その手続きを示した．

4.2.2 二値分類器の構築と適用

特徴量を設計した後，攻撃者は，D_X とは異なるデータセット $D_O = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_{tr}}\}$ を用いて学習を行い，スコア関数としての二値分類器を構築する．ここで，$\boldsymbol{z}_i = \{z_{i,j}\}_{j \in J_X \cup J_Y}$ $(i \in \{1, \cdots, N_{tr}\})$ である．$\delta_{i,j}$ をクロネッカーのデルタとし，$l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ を適当な損失関数とする．$\mathbb{R}^+$ は非負の実数集合を指す．このとき，攻撃者は以下のように期待損失を最小化する二値分類器 f^* の構築を試みるとする:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}[\ l(f(\{z_{i_y, j_y}\}_{j_y \in J_Y}, \{z_{i_x, j}\}_{j \in J_X}), \delta_{i_y, i_x})\ ]$$

ここで，$\mathcal{F}$ は，値域を $[0, 1]$ に持つ二値分類器の集合とする．しかし，一般には上記の最適化問題を解くことは困難であるため，経験誤差と正則化項の和を最小化する解を求めることが多い．本研究では，そうした手続きを行う決定木ベースの手法の一つである XGBoost という手法 [10] を用いた．この手法は，機械学習の様々な問題において最高精度を達成することが知られており，二値分類問題もその例外ではないため，プライバシリスクの評価にも有効であると考えられる．なお，損失関数や最適化の詳細については元論文 [10] を参照していただきたい．

学習により攻撃者が構築した関数を f_{bc} とする．その後，攻撃者は以下のように f_{bc} を適用し，攻撃対象の個人のレコードの予測を行うとする:

$$\hat{i} = \arg\max_{i \in \{1, 2, \cdots, n\}} f_{bc}(\boldsymbol{x}_i, \boldsymbol{y})$$

5. 実験

5.1 データセットと計算環境

本研究では，HapMap というプロジェクトにより公開されているゲノムデータを使用して実験を行った [*1]．このプロジェクトは，ヒトの各染色体について，個人の SNP を diplotype genotypes の形式で保存したデータセットをウェブ上に公開している．我々は，CEU (Utah Residents with Northern and Western European Ancestry) についての染色体 1 のデータセットを選んだ．データセットは定期的に更新されているが，本研究では 2009 年 2 月 phese II+III として公開されている genotypes データセットを使用した．このデータセットには，314,024 個の SNP についての情報が 174 人分含まれている．

さらに，HapMap プロジェクトは，genotypes データセット以外にも連鎖不平衡についてのデータセットを公開している．本研究では，genotypes データセットとの整合性をとれるよう，2009 年 4 月のデータセットを使用した．このデータセットには，44,332,127 個の SNP ペアについての r^2 統計量などが含まれている．

本研究の実験は Ubuntu 14.04（AWS EC2) 上で行った．実装は Python 2.7.11 で行い，XGBoost の部分には公開されているパッケージを用いた [*2]．なお，学習率は 0.3 とし，その他のハイパーパラメータは指定せず計算を行った．

[*1] http://www.ncbi.nlm.nih.gov (2016 年 7 月 31 日現在)
[*2] https://github.com/dmlc/xgboost（バージョンは 0.4 を使用）

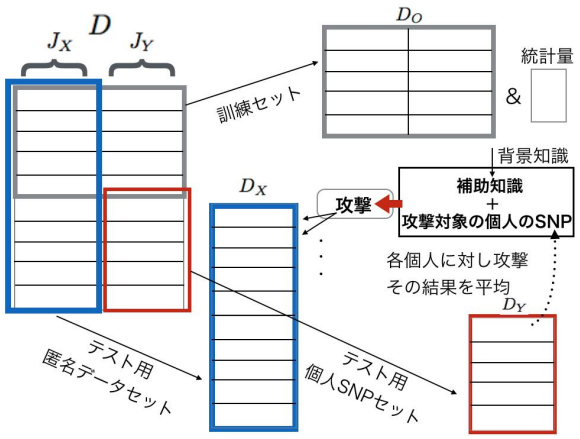

図 5: 実験の概要．D_Y は学習に使用されていないため，脱匿名化攻撃の Recall を評価する際には，D_X と D_O を異なるデータセットと見なして学習とテストを行っても過学習を回避できる．

5.2 実験設計

本研究では，様々な強さを持った攻撃者を考慮するために，様々な J_X と J_Y に対して実験を行う．J_X と J_Y のサンプリング方法を 5.2.1 節で説明する．また，各 J_X と J_Y の組に対して 5.2.2 節で示すように学習とテストを行う．

5.2.1 J_X と J_Y のサンプリング

公開されている HapMap の genotypes データセットは非常に多くの SNP を含んでいるため，すべての SNP を扱って実験を行うことは現実的ではない．そこで，本研究の提案手法を評価するための J_X と J_Y のサンプリング手法を考える．まず，ある閾値 $r^2_{\text{thres}} \in \{0.1, 0.4, 0.7\}$ を設定する．次に，アルゴリズム 1 によって J_X と J_Y を生成する．このアルゴリズムは，r^2_{thres}，M，Pair-LD の三つを入力とする．四行目の手続きによって，J_X と J_Y に振り分けられる各座位が，r^2_{thres} の値によって与えられる制約を満たすペアの中からサンプリングされている．この制約により，アルゴリズムの出力である J_X と J_Y の相関関係が以下の意味で制御されていることが分かる: J_X に含まれる任意の座位 j は，少なくとも一つ以上の $j_y \in J_Y$ に対して，$r^2(j, j_y)$ が四行目の制約を満たしており，かつ J_Y に含まれる任意の座位 j_y は，少なくとも一つ以上の $j \in J_X$ に対して，$r^2(j, j_y)$ が四行目の制約を満たしている．なお，実験において r^2_{thres} を様々な値に設定することにより，J_X と J_Y の多様性を J_X と J_Y の間の相関関係の多様性によって与えることが可能になる．特に，r^2_{thres} が大きいほど攻撃者は強くなるという仮説が立てられる．我々はこの仮説を 5.3 節において検証した．

5.2.2 学習とテスト

アルゴリズム 1 により (J_X, J_Y) が与えられた後，アルゴリズム 2 により (J_X, J_Y) についての学習とテストを行う．まず，$J_X \cup J_Y$ の座位の SNP をデータベースから全員分取得する．取得したデータセットを D と表記する．次に，図 5 のように，D を訓練セット D_O とテストセットにランダムに分割する．なお，データの分割にあたっては，一般にはテストセットとして訓練セットに含まれない J_X と J_Y からなるデータのみを用いて評価を行うことが多い．しかし，本研究では実験に使用できるサンプルサイズが小さいため，そうした実験設定ではランダムな予測による Recall が高くなる恐れがある．実際，本研究の実験では，訓練セットとテストセットを 4 : 1 の比で分割した．このとき，D_Y に含まれる人の数は 34 人であるため，従来の設定によるテストセットでは，ランダムな予測による Recall の期待値は $1/34 \simeq 2.94\%$ となってしまう．一般に，ランダムな予測による Recall が低いような実験設定の下で提案手法が高い Recall を達成するとき，提案手法の有効性がより強く示されることになる．そこで，我々は図 5 に示したような D_X と D_Y をテストセットとして設定した．このとき，D_X に含まれる個人は 174 人であったため，ランダムな予測による Recall の期待値は $1/174 \simeq 0.57\%$ となり，

アルゴリズム 1 実験に使用する SNP の座位の生成

function Generator(r^2_{thres}, M, Pair-LD)
Input: r^2_{thres} : サンプリングの閾値．$M(= |J_X| = |J_Y|)$: 攻撃に使用される SNP の数．Pair-LD: HapMap により公開された LD の情報を含むデータセット．
Output: (J_X, J_Y)

1: $num \leftarrow 1$
2: $J_X \leftarrow [\,], J_Y \leftarrow [\,]$
3: **while** $num \leq M$ **do**
4: $(j, j_y) \sim$ Pair-LD (uniformly sampled)
s.t. $r^2_{\text{thres}} \leq r^2(j, j_y) \leq r^2_{\text{thres}} + 0.3$
5: **if** $j \notin J_X \cup J_Y$ **and** $j_y \notin J_X \cup J_Y$ **then**
6: **append** $j \rightarrow J_X$, $j_y \rightarrow J_Y$
7: $num \leftarrow num + 1$
8: **end if**
9: **end while**
10: **return** (J_X, J_Y)

アルゴリズム 2 学習とテスト

function Train-and-Test(D, D_O, D_X, D_Y Pair-LD)
Input: D: 実験で使用する全体のデータセット．D_O: 訓練セット．D_X: テスト用匿名データセット．D_Y: テスト用個人 SNP セット．Pair-LD: アルゴリズム 1 と同じデータセット．
Output: *Recall*: True Positive Rate.

1: $f \leftarrow D_O$ を用いて学習されるスコア関数．
2: $Recall \leftarrow 0$
3: $n \leftarrow D_X$ に含まれるレコード数
4: $n_y \leftarrow D_Y$ に含まれるレコード数
5: $\boldsymbol{x}_i \leftarrow D_X$ 内の i 番目の個人の SNP $(i \in \{1, \cdots, n\})$
6: **for all** $i_{target} \in \{1, \cdots, n_y\}$ **do**
7: $\boldsymbol{y} \leftarrow D_Y$ 内の i_{target} 番目の個人の SNP
8: $\hat{i}_{target} \leftarrow \arg\max_{i \in \{1, \cdots, n\}} f(\boldsymbol{x}_i, \boldsymbol{y})$
9: **if** $\hat{i}_{target} == i_{target}$ **then**
10: $Recall \leftarrow Recall + 1$
11: **end if**
12: **end for**
13: $Recall \leftarrow Recall / n_y$
14: **return** *Recall*

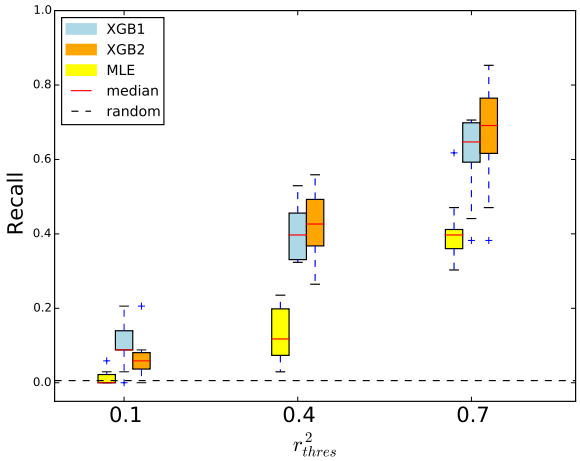

図 6: 実験結果．$|J_X| = |J_Y| = 20$ の場合の Recall

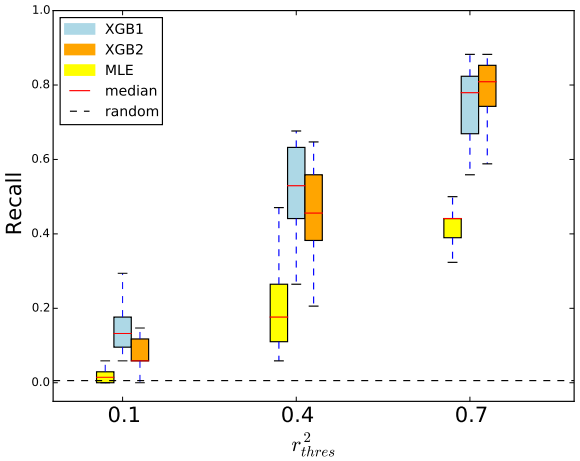

図 7: 実験結果．$|J_X| = |J_Y| = 30$ の場合の Recall

その値は従来の設定と比べ 1/5 倍ほど低くなる．

以上のような実験設定の下で，攻撃者は，訓練セット D_O を用いて 4 章で説明したようにスコア関数 f を構築する．なお，このとき，D_O に加えて，Pair-LD の情報も攻撃者の補助知識として使用することができるとする．

学習後，攻撃者は D_Y から攻撃対象の個人を一人ずつ順に選択し，選択した個人と D_X 内の各レコードに対して図 5 のように関数 f を適用して脱匿名化攻撃を行う．D_Y 全体で結果を平均して Recall を算出する．

5.3 結果と考察

M の値として $\{20, 30\}$ の二通りを考え，それぞれの M に対して，r^2_{thres} の値として $\{0.1, 0.4, 0.7\}$ の三通りを設定して実験を行った．それぞれの場合について，J_X と J_Y のサンプリングがランダム性を持つことを考慮し，アルゴリズム 1, 2 を 10 回ずつ順に適用することで実験結果の信頼性を確保した．$M = 20$ の場合の結果を箱ひげ図として出力したものを図 6 に示す．図の縦軸は Recall であり，横軸は r^2_{thres} である．MLE は最尤推定を使用した結果である．XGB1 は特徴量設計の前処理において次元削減を行った二値分類を使用した結果であり，XGB2 は次元削減を行わない二値分類を使用した結果である．箱の上辺は 10 回の実験における第 3 四分位点であり，下辺は第 1 四分位点に相当する．箱の中の赤線は各実験での中央値を示す．箱の上に書かれた短い横線は，“最大値と，第 3 四分位点 +1.5（第 3 四分位点 − 第 1 四分位点）” のうち小さい方の値を示す．箱の下に書かれた同様の線は，“最小値と，第 1 四分位点 −1.5（第 3 四分位点 − 第 1 四分位点）” のうち大きい方の値を示す．また，図全体を横切って水平に引かれた黒の破線はランダムな予測結果の期待値を示す．図 6 から以下のことが分かった．

まず，サンプリングの閾値は攻撃者の強さを制御しており，r^2_{thres} が大きいほど攻撃者は強くなるという仮説が実験的に確かめられた．

次に，SNP 間の相関情報は脱匿名化攻撃に有効な背景知識であることが確かめられた．この考察は，本研究の提案手法が攻撃者の背景知識として SNP 間の相関情報を設定しており，かつ図 6 において全ての提案手法がランダムな予測と比べて高い性能を示していることから導かれる．特に，攻撃者が強い（$r^2_{\mathrm{thres}} = 0.7$ で，かつ二値分類を使用した攻撃を行う）とき，Recall は約 70% であり，ランダムな予測の約 122 倍の精度で脱匿名化が達成される．

さらに，二値分類は最尤推定よりも高精度な脱匿名化リスクの評価を行えるということが確かめられた．この仮説はすべての r^2_{thres} 値において確かめられた．

なお，次元削減が精度を高めるという仮説は，$r^2_{\mathrm{thres}} = 0.1$ の場合において確かめられた一方で，他の場合では確かめられなかった．この理由としては，$r^2_{\mathrm{thres}} = 0.1$ の場合では次元削減により有効な特徴量の選別が容易になった一方で，他の場合では次元削減をせずとも有効な特徴量を選別できていたことが考えられる．さらに，$r^2_{\mathrm{thres}} = 0.4$ の場合にも次元削減が精度を高めているという点を除いて，同様の傾向が $M = 30$ の場合の結果を示した図 7 に関しても言える．

また，$M = 30$ の場合は $M = 20$ の場合と比較して脱匿名化の精度が高くなっていることが示された．この結果は，$M = 30$ の場合の方が攻撃者の知識が多いということを反映していると考えられる．

6. 関連研究

6.1 症例群に個人が存在しているか否かの推定攻撃

Homer らは，攻撃対象の個人のプロフィール情報を用いて，公開されたゲノム研究の症例群に攻撃対象の個人が含まれているか否かを推定する攻撃を提案した [4]．Wang ら

はこの研究を進め，SNP 間のペアワイズの相関情報を利用して，より強力な攻撃を提案した [5]．Wang らの論文では，攻撃対象の個人の SNP を特定する攻撃も行われていたが，本研究が対象としている匿名ゲノムデータベースに対する脱匿名化攻撃に関する評価は行われていない．

6.2 匿名データセットに存在する個人に対する脱匿名化攻撃

Lin らは，攻撃対象の個人についての explicit knowledge として約 100 個の SNP を観測した攻撃者は，高い確率で匿名ゲノムデータベースに対する脱匿名化攻撃を行えることを示した [3]．Humbert らは，攻撃対象の個人の形質情報を利用した匿名ゲノムデータベースに対する脱匿名化攻撃を提案した [1]．Humbert らの研究では，提案手法の攻撃精度は数%程度であり，ランダムな予測と比較して約 4 倍の精度であったと報告されている．ただし，彼らの研究で使用されたデータや想定されている攻撃者の強さは本研究とは異なっているため，数値そのものを単純に比較して精度の優劣をつけることは難しい．また，Humbert らの研究では訓練データとテストデータの分割が行われていないため，過学習の問題を回避した精度評価が行われているとは言い切れないが，5 章で述べたように本研究ではその問題を回避した精度評価を行っている．ゲノムデータベース以外を扱った研究としては，Narayanan と Shmatikov が，Netflix という映画評価データベースに対して，関連するデータベースの情報を用いた脱匿名化攻撃を提案した [6]．彼らは，似た設定の下で，オンライン SNS のユーザーアカウントに対する脱匿名化攻撃も提案した [7]．Korayem と Crandall は，オンライン SNS のユーザーアカウントに対する二値分類器を用いた脱匿名化攻撃を提案した [8]．我々が知る限り，二値分類器を用いた脱匿名化攻撃の研究に関しては，本研究で扱ったような特徴ドメインの性質や特徴ベクトルの設計を考慮したものは存在しない．

6.3 個人の SNP の推定攻撃

Nyholt らは，James Watson という人物の ApoE ゲノムを，本人が公開している他の SNP を基に推定する攻撃を行った [9]．また，Samani らは，個人の SNP を，その個人が公開している他の SNP を基に推定する攻撃を行った [2]．彼らは，SNP 間の低次相関や高次相関を利用した様々な攻撃モデルを提案し，推定した SNP と真の SNP との絶対誤差を攻撃精度の指標として評価した．しかし，彼らの研究は脱匿名化攻撃を扱ってはおらず，本研究とは問題設定が異なっている．

7. おわりに

我々は，脱匿名化攻撃における問題設定と攻撃手法の提案を通して，これまで考えられていなかった攻撃者に対するプライバシリスクを評価することを目的とした研究を行った．問題設定の面では，既存の脱匿名化攻撃の枠組みを整理し，攻撃者が背景知識として implicit な SNP を持つ場合の攻撃モデルを提案した．また，手法の面では，既存の最尤推定を用いた攻撃手法を改良するとともに，より高精度な脱匿名化攻撃を可能にする二値分類を用いた手法を提案した．実データを用いた実験により，様々な強さの攻撃者を考慮して，それぞれの場合におけるプライバシリスクを評価した．結果として，全ての提案手法がランダムな予測と比べて高い精度で脱匿名化攻撃を行うことが示された．今後は，より現実的あるいは強力な問題設定の下での攻撃モデルを考えつつ，こうしたプライバシリスクの評価を基にして，適切なプライバシ保護のモデルについても研究を進めていきたい．

参考文献

[1] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux,“ De-anonymizing genomic databases using phenotypic traits, ” Proceedings on Privacy Enhancing Technologies, vol. 2015, no. 2, pp. 99-114, 2015.

[2] S. S. Samani, Z. Huang, E. Ayday, M. Elliot, J. Fellay, J.-P. Hubaux, and Z. Kutalik. Quantifying genomic privacy via inference attack with high-order SNV correlations. In Workshop on Genome Privacy. 32-40, 2015.

[3] Z. Lin, A. B. Owen, and R. B. Altman. Genomic research and human subject privacy. Science, 305(5681):183, Jul 2004.

[4] N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 4, Aug. 2008.

[5] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. CCS ’ 09: Proc. of the 16th ACM Conf. on Computer and Communications Security, pages 534-544, 2009.

[6] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In SP ’ 08: Proc. of the 29th IEEE Symp. on Security and Privacy, pages 111-125, 2008.

[7] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In SP ’ 09: Proc. of the 30th IEEE Symp. on Security and Privacy, pages 173-187, 2009.

[8] M. Korayem and D. J. Crandall. De-anonymizing users across heterogeneous social computing platforms. In Proceedings of 7th International AAAI Conference on Weblogs and Social Media (ICWSM), 2013.

[9] D. R. Nyholt, C.-E. Yu, and P. M. Visscher, “ On Jim Watson’s APOE status: genetic information is hard to hide, ” European Journal of Human Genetics, vol. 17, no. 2, p. 147, 2009.

[10] T. Chen. and C. Guestrin, “ XGBoost: A Scalable Tree Boosting System, ” arXiv:1603.02754, 2016.