プライバシーを保護するカウント演算の多値属性分類への適用について

On Applying Privacy Preserving Count Aggregate Queries to k-Classification

高見澤 秀久 ♥ 有次 正義 ♦
Hidehisa TAKAMIZAWA Masayoshi ARITSUGI

プライバシーを保護しながらデータを効果的に処理することは重要な課題である．本稿では，プライバシー保護のために摂動されたテーブルから，目的属性が 3 値以上の決定木を構築するために必要なカウント演算結果を再構築する手法を提案する．目的属性が 3 値以上の場合，従来手法では目的属性の各値の演算結果をそれぞれ独立に再構築しなければならない．そこで，本稿では従来手法を拡張し，目的属性の各値の演算結果を一括して再構築する手法を提案する．

It is important to process data effectively while preserving privacy. In this paper, we propose a reconstruction technique of count aggregate queries, which are necessary for building a decision tree, from a perturbed table in cases where a target attribute is more than binary. In the conventional technique, we must reconstruct the results of target values from those of each value calculated independently when a decision tree has a non-binary target attribute. In this paper, we borrow and extend the conventional technique to reconstruct the results of target values at once.

1. はじめに

近年，プライバシー保護データマイニングの研究が盛んに行われている [1, 2, 3]．本研究では，サーバと，そのサーバに接続された n 個のクライアント $Client_1$，$Client_2$，$\cdots$，$Client_n$ により構成されるモデルを考える．各クライアントは $m+1$ 個の属性 $Attr_0$，$Attr_1$，$Attr_2$，$\cdots$，$Attr_m$ からなるレコードをそれぞれ 1 つずつもっている．ここで，属性 $Attr_0$ は決定木分類における目的属性でカテゴリ属性である．他の属性 $Attr_1$，$\cdots$，$Attr_m$ は説明属性である．各クライアントは自身がもつレコードのプライバシーを保護するために，摂動（perturbation）されたレコードをサーバに送信する．サーバは，すべてのクライアントから受信した n 個の摂動されたレコードから，テーブル T' を構成する．すなわち，サーバは各クライアントにおいて摂動されたレコードの値は知ることができるが，各クライアントが所持する元のレコードの値を知ることはできない．したがって，このモデルでは摂動されていない元のテーブル T が構成されることはない．ただし，以降は説明を簡略化するため，テーブル T の各レコードを独立に摂動することで，テーブル T' を構成するものと考える．

本研究の目的は，摂動されたテーブルから，目的属性が 3 値以上の決定木を構築するために必要なカウント演算結果を再構築することである．本研究と同じモデルで摂動されたテーブルからカウント演算結果を再構築する手法は文献 [3] においても提案されている．文献 [3] では，ある一定の確率で同じ属性の定義域内の別のランダムな値に置き換える方法により摂動されたテーブルから，決定木を構築するために必要なカウント演算結果を再構築する手法を提案している．

しかしながら，文献 [3] で提案されている手法を用いて，摂動されたテーブルから目的属性が 3 値以上である決定木を構築するために必要なカウント演算結果を再構築する場合，目的属性の各値の演算結果をそれぞれ独立に再構築しなければならない．そこで，本稿では文献 [3] で提案されている再構築手法を拡張して，目的属性の各値の演算結果を一括して再構築する手法を提案する．

2. 準備

2.1 関連研究

本研究におけるプライバシー保護モデルは，文献 [1, 3] で扱われている．文献 [1] では，各クライアントがもつレコードに対して，属性ごとに独立な乱数をノイズとして加算することでプライバシー保護を行う．そして，摂動されたテーブルから各属性の分布を再構築する．しかしながら，文献 [1] では，元の分布の再構築は各属性で独立に行わなければならないという問題がある．文献 [3] では，ある一定の確率で同じ属性の定義域内の別のランダムな値に置き換える手法により，各クライアントがもつレコードを摂動する．そして，摂動されたテーブルから，決定木を構築するために必要なカウント演算結果を再構築する．本研究は文献 [3] を基に考える．以下では，文献 [3] で提案されている摂動手法と再構築手法について説明する．

2.2 プライバシー保護 OLAP

■維持 - 置換摂動 文献 [3] では，各レコードを摂動するために維持 - 置換摂動（Retention Replacement Perturbation）を用いている．維持 - 置換摂動は，レコードの各属性値を，属性ごとに共通なある一定の確率で同じ属性の定義域内の別の値に置換する手法である．すなわち，この手法では摂動後も元の値がある一定の確率で維持される．この確率を維持確率（retention probability）として属性ごとに設定する．ここでは，ある属性 $Attr_j$ における維持確率を $rp_j (0 \leq rp_j \leq 1)$ と表現する．維持

♥ 学生会員 群馬大学大学院工学研究科博士後期課程 takamiza@dbms.cs.gunma-u.ac.jp
♦ 正会員 熊本大学大学院自然科学研究科 aritsugi@cs.kumamoto-u.ac.jp

確率が小さい，つまり元の値が維持されずに他の値に置換される確率が高くなるほど摂動の割合も高くなり，プライバシーが保護されることになる．

ここで，t_{ij} は元のテーブル T の i 番目のレコードにおける j 番目の属性値，そして t'_{ij} は摂動されたテーブル T' の i 番目のレコードにおける j 番目の属性値とすれば，t'_{ij} の値は確率 rp_j で t_{ij} に，そして確率 $1-rp_j$ で定義域内の他の値に置換される．なお，本研究におけるモデルでは，各クライアントはすべての属性の維持確率と定義域を知っているものとし，それにより各クライアントは自身がもつレコードを独立に摂動することができる．

■反復ベイズ手法 続いて，文献 [3] で提案されている，維持 - 置換摂動により摂動されたテーブル T' からカウント演算結果を再構築する手法である，反復ベイズ手法（Iterative Bayesian Technique）について説明する．ここでは，テーブル T における k 個の属性に対してカウント演算 $COUNT(P_1 \wedge P_2 \wedge \ldots \wedge P_k)$ を実行した結果を再構築することを考える．なお，P_j はテーブル T における j 番目の属性に対する条件式であり，例えば数値属性の場合は $25 \leq age \leq 40$，カテゴリ属性の場合は $risk = high$ といった条件式となる．

まず，元のテーブルにおけるカウント演算結果を表すベクトル $\boldsymbol{x} = (x_0, x_1, ..., x_t)$ と，摂動されたテーブルにおけるカウント演算結果を表すベクトル $\boldsymbol{y} = (y_0, y_1, ..., y_t)$ を以下のように定義する．

$$\begin{cases} x_i = COUNT(\wedge_{r=1}^{k} Q(r, bit(i,r))) \\ \qquad\qquad \text{in } T, \text{ for } 0 \leq i \leq t \\ y_i = COUNT(\wedge_{r=1}^{k} Q(r, bit(i,r))) \\ \qquad\qquad \text{in } T', \text{ for } 0 \leq i \leq t \end{cases} \tag{1}$$

ここで，$t = 2^k - 1$ である．また，$Q(r,i)$ は r 番目の属性に対する条件式であり，i の値が 0 ならば $\neg P_r$ に，そして i の値が 1 ならば P_r となる．$bit(i,r)$ は，整数 i を k ビットの二進数で表現した値の左から r 番目のビットである．そして，あるレコードが条件 $\wedge_{r=1}^{k} Q(r, bit(i,r))$ を満たすとき，そのレコードの状態を i とする．

次に，元のテーブル T での状態が p であったレコードが，摂動されたテーブル T' での状態が q となる遷移確率 a_{pq} は，以下の式により計算することができる．

$$a_{pq} = \prod_{r=1}^{k} \left\{ (1 - rp_r) \cdot R_{r,bit(q,r)} + rp_r \cdot \delta_{(bit(p,r),bit(q,r))} \right\} \tag{2}$$

ここで，$R_{r,i}$ は，i の値が 0 ならば $1 - b_r$，i の値が 1 ならば b_r と定義する．b_r は条件式の範囲を表すものであり，以下の式により計算する．

$$b_r = \begin{cases} \dfrac{high_r - low_r}{max_r - min_r}, & (Attr_r \text{ が数値属性の場合}) \\ \dfrac{1}{c_r}, & (Attr_r \text{ がカテゴリ属性の場合}) \end{cases} \tag{3}$$

属性 $Attr_r$ が数値属性の場合，$high_r$，low_r は，それぞれ属性 $Attr_r$ に対する条件式 P_r の上限値と下限値であり，max_r，min_r は，それぞれ属性 $Attr_r$ の定義域の上限値と下限値である．属性 $Attr_r$ がカテゴリ属性の場合，c_r は属性 $Attr_r$ の属性値の個数である．$\delta_{(i,j)}$ は i と j の値が等しければ 1 を，そうでなければ 0 とする．

続いて，これらの式を用いて，摂動されたテーブル T' におけるカウント演算結果 $\boldsymbol{y}$ から元のテーブル T におけるカウント演算結果 $\boldsymbol{x}$ を再構築する手順を示す．まず，テーブル T における n 個の各レコードの状態をそれぞれ $U_1, U_2, ..., U_n$，テーブル T' における n 個の各レコードの状態をそれぞれ $V_1, V_2, ..., V_n$ とする．すなわち，$0 \leq p, q \leq 2^k - 1$ および $1 \leq i \leq n$ に対して，$P(U_i = p) = x_p/n$，$P(V_i = q) = y_q/n$ となる．また，$P(U_i = p)$ は以下の式で表すことができる．

$$P(U_i = p) = \sum_{q=0}^{t} P(V_i = q) P(U_i = p | V_i = q) \tag{4}$$

ここで，$P(U_i = p | V_i = q)$ は，ベイズの定理を用いて以下の式により表すことができる．

$$\begin{aligned} P(U_i = p | V_i = q) &= \frac{P(V_i = q | U_i = p) P(U_i = p)}{P(V_i = q)} \\ &= \frac{P(V_i = q | U_i = p) P(U_i = p)}{\sum_{r=0}^{t} P(V_i = q | U_i = r) P(U_i = r)} \\ &= \frac{a_{pq} x_p}{\sum_{r=0}^{t} a_{rq} x_r} \end{aligned} \tag{5}$$

そして，式 (4) に，式 (5) を代入することで，以下の式を求めることができる．

$$x_p^{T+1} = \sum_{q=0}^{t} y_q \frac{a_{pq} x_p^T}{\sum_{r=0}^{t} a_{rq} x_r^T} \tag{6}$$

ここで，$\boldsymbol{x}^T$ は，T 番目の繰返しを，$\boldsymbol{x}^{T+1}$ は，$T+1$ 番目の繰返しをそれぞれ示しており，初期値を $\boldsymbol{x}^0 = \boldsymbol{y}$ として，連続する $\boldsymbol{x}$ の違いが少なくなるまで，繰返し $\boldsymbol{x}$ を計算する．

3. 摂動されたテーブルからの決定木の構築

決定木 [4, 5, 6] における最適な分割点を決定するための評価指標としては，一般に相互情報量（mutual information），ジニ係数（gini index），χ^2（chi square）値等が用いられる [6, 7, 8]．いずれの評価関数も，分割前後の各属性値の個数の関数として表現することができるため，最適な分割を探索する際にはカウント演算結果を定数として扱うことができる．

3.1 2 値属性分類のためのカウント演算

まず，反復ベイズ手法を用いて，目的属性が 2 値 C_0，C_1 の決定木を構築するために必要なカウント演算結果を摂動されたテーブルから再構築する手順を説明する．

目的属性が 2 値の場合は，$C_1 = \neg C_0$ である．したがって，条件 P_1 により分割する場合，目的属性が 2 値の決定木を構築するために必要なカウント演算は反復ベイズ手法により，一度に計算することができる．

3.2 多値属性分類のためのカウント演算

続いて，反復ベイズ手法を用いて，目的属性が 3 値以上の決定木を構築するために必要なカウント演算結果を摂動されたテーブルから再構築する手順を説明する．目的属性が 3 値以上の場合は，属性値ごとのカウント演算結果が必要となる．例えば，目的属性が 3 値 C_0, C_1, C_2 である集合を条件 P_1 により分割する場合，表 1 のカウント演算が必要となる．

表 1 3 値属性分類のためのカウント演算
Table 1 Count aggregate queries for a 3-values target attribute.

目的属性	条件 P_1 で分割する場合のカウント演算
C_0	$COUNT(P_0^0 \wedge \neg P_1)$
	$COUNT(P_0^0 \wedge P_1)$
C_1	$COUNT(P_0^1 \wedge \neg P_1)$
	$COUNT(P_0^1 \wedge P_1)$
C_2	$COUNT(P_0^2 \wedge \neg P_1)$
	$COUNT(P_0^2 \wedge P_1)$

ここで，P_0^i は目的属性の属性値が C_i であることを表す条件式である．また，従来手法により一回の処理で再構築できるカウント演算を表 2 に示す．なお，$\boldsymbol{x}^i$ を，目的属性の属性値が C_i であるか否かの条件 P_0^i または $\neg P_0^i$ をもつカウント演算の結果を表すベクトルとする．

表 2 従来手法によって再構築可能なカウント演算
Table 2 Reconstructable count aggregate queries by the conventional technique.

目的属性	条件 P_1 で分割する場合のカウント演算	$\boldsymbol{x}^i$
C_i	$COUNT(\neg P_0^i \wedge \neg P_1)$	x_{00}^i
	$COUNT(\neg P_0^i \wedge P_1)$	x_{01}^i
	$COUNT(P_0^i \wedge \neg P_1)$	x_{10}^i
	$COUNT(P_0^i \wedge P_1)$	x_{11}^i

そのため，従来手法では，目的属性の各値のカウント演算結果 $\boldsymbol{x}^1$，$\boldsymbol{x}^2$，$\boldsymbol{x}^3$ をそれぞれ独立に再構築しなければ，表 1 に示すすべてのカウント演算，すなわち，$COUNT(P_0^i \wedge \cdots)$ の結果を得ることができない．

4. 提案手法

前述の通り，摂動されたテーブルから，目的属性が 3 値以上の決定木を構築するために必要なカウント演算結果を従来手法を用いて再構築する場合，目的属性の各値においてそれぞれ独立に再構築処理を行わなくてはならない．そこで，本節では，目的属性の各値のカウント演算結果を一括して再構築する手法を提案する．

4.1 レコードの状態の再定義

まず，2. で定義したレコードの状態を再定義する．従来手法では，各属性の状態を示すビットにより，条件 P_i，$\neg P_i$ を満たすかどうかをそれぞれ 1，0 として表現していた．目的属性が 2 値の場合には，あるレコードの目的属性が「一方の属性値をもつ／もたない」として，目的属性の状態を 1 ビットで表現することができた．しかしながら，目的属性が 3 値以上の場合には，目的属性の状態を 1 ビットで表現することはできない．そこで，本研究では目的属性の状態をビット列として表現する．

目的属性の属性値の個数を c とすれば，目的属性の状態を表すために必要となるビット数は，$\lceil \log_2 c \rceil$ により計算できる．したがって，レコードの状態は，目的属性の状態を表す長さ $\lceil \log_2 c \rceil$ のビット列と，m 個の説明属性の状態を表す長さ m のビット列を連結したものとなる．

例として，3 値 C_0, C_1, C_2 を属性値としてもつ目的属性と，1 つの説明属性からなるレコードが取り得る状態を，その状態に対する条件式と対応付けて表 3 に示す．ここで，P_0^i は目的属性の属性値が C_i であることを表す条件式，そして P_1 は説明属性に対する条件式である．

表 3 提案手法におけるレコードの状態
Table 3 Status of a record in our proposal.

状態	条件式	目的属性の値
000	$P_0^0 \wedge \neg P_1$	C_0
001	$P_0^0 \wedge P_1$	
010	$P_0^1 \wedge \neg P_1$	C_1
011	$P_0^1 \wedge P_1$	
100	$P_0^2 \wedge \neg P_1$	C_2
101	$P_0^2 \wedge P_1$	

ここでは，目的属性が 3 値であるため，目的属性の状態は $\lceil \log_2 3 \rceil = 2$ ビットで表現することができる．また，説明属性の状態は 1 ビットで表現することができるため，レコードの状態は $2 + 1 = 3$ ビットで表現することができる．

4.2 カウント演算の定式化

次に，4.1 で定義した各状態に対するカウント演算を定式化する．反復ベイズ手法と同様に，元のテーブルにおけるカウント演算結果を表すベクトル $\boldsymbol{x} = (x_0, x_1, ..., x_t)$ と，摂動されたテーブルにおけるカウント演算結果を表すベクトル $\boldsymbol{y} = (y_0, y_1, ..., y_t)$ を以下のように定義する．

$$\begin{cases} x_i = COUNT(P_0^{left(i, \lceil \log_2 c \rceil)} \wedge_{r=1}^{k} Q(r, bit(i,r))) \\ \qquad \text{in } T, \text{ for } 0 \le i \le t' \\ y_i = COUNT(P_0^{left(i, \lceil \log_2 c \rceil)} \wedge_{r=1}^{k} Q(r, bit(i,r))) \\ \qquad \text{in } T', \text{ for } 0 \le i \le t' \end{cases} \quad (7)$$

ここで $t' = c \cdot 2^k - 1$ である．また，$left(i, r)$ を i の左から r ビットの整数表現とする．したがって，$left(i, \lceil \log_2 c \rceil)$ は，値 i の左から $\lceil \log_2 c \rceil$ ビット，すなわち目的属性の属性値を表して

いるため，式 $P_0^{left(i,\lceil \log_2 c \rceil)}$ は $left(i, \lceil \log_2 c \rceil)$ によって表される属性値を選択するための条件式となる．例えば，目的属性が 3 値，$i = 010_{(2)}$ の場合，$P_0^{left(010_{(2)},2)} = P_0^{01_{(2)}} = P_0^1$ となる．なお，$Q(r,i)$ および $bit(i,j)$ の定義は反復ベイズ手法と同様である．

例として，3 値の目的属性と 1 つの説明属性からなるテーブルを考えた場合，$\boldsymbol{x}$ および $\boldsymbol{y}$ に対応するカウント演算を表 4 に示す．

表 4　提案手法によって再構築可能なカウント演算
Table 4　Reconstructable count aggregate queries by the proposed technique.

カウント演算	$\boldsymbol{x}$	$\boldsymbol{y}$	目的属性
$COUNT(P_0^0 \wedge \neg P_1)$	x_{000}	y_{000}	C_0
$COUNT(P_0^0 \wedge P_1)$	x_{001}	y_{001}	
$COUNT(P_0^1 \wedge \neg P_1)$	x_{010}	y_{010}	C_1
$COUNT(P_0^1 \wedge P_1)$	x_{011}	y_{011}	
$COUNT(P_0^2 \wedge \neg P_1)$	x_{100}	y_{100}	C_2
$COUNT(P_0^2 \wedge P_1)$	x_{101}	y_{101}	

ここでは，目的属性の状態は，$left(i, \lceil \log_2 3 \rceil) = 2$ ビットにより表すことができるため，i の左から 2 ビットは目的属性，そして残りの 1 ビットは説明属性の状態として表される．

4.3　遷移確率の再定義

続いて，状態 p のレコードが状態 q となる遷移確率 a'_{pq} を以下に示す．

$$a'_{pq} = \left\{(1 - rp_0) \cdot b_0 + rp_0 \cdot \delta_{(left(p,\lceil \log_2 c \rceil), left(q,\lceil \log_2 c \rceil))}\right\} \cdot \prod_{r=1}^{k} \left\{(1 - rp_r) \cdot R_{r,bit(q,r)} + rp_r \cdot \delta_{(bit(p,r),bit(q,r))}\right\} \tag{8}$$

説明属性においては，条件 P_r と $\neg P_r$ の範囲が一致するとは限らないので，置換された値が元の状態ではなくなる確率 $1 - b_r$ と，元の状態と同じになる確率 b_r が一致するわけではない．一方，目的属性においては，条件 P_0^i の範囲はすべて等しくなるため，置換された値が各属性値になる確率は均一に $b_0 = 1/c$ である．また，説明属性の状態の変化は 1 ビットを比較することで判別できるが，目的属性の状態の変化は目的属性の状態を表す長さ $left(i, \lceil \log_2 c \rceil)$ のビット列を比較する必要がある．

4.4　再構築の計算式

再構築の計算式は，従来手法の式 (6) の遷移確率 a_{pq} を式 (8) の遷移確率 a'_{pq} に置き換える．すなわち，以下の式により表される．

$$x_p^{T+1} = \sum_{q=0}^{t'} y_q \frac{a'_{pq} x_p^T}{\sum_{r=0}^{t'} a'_{rq} x_r^T} \tag{9}$$

そして，反復ベイズ手法と同様に，$\boldsymbol{x}$ を繰返し計算すればよい．

5.　おわりに

本稿では，維持 - 置換摂動によって摂動されたテーブルから，目的属性が 3 値以上の決定木を構築するために必要なカウント演算結果を再構築する手法を提案した．従来手法を用いて目的属性が 3 値以上のカウント演算結果を摂動されたテーブルから再構築する場合，目的属性の各属性値で独立に再構築処理を行う必要があった．そこで，本稿では目的属性の各属性値を一括して再構築する手法を提案した．

[謝辞]

本研究の一部は，科学研究費補助金基盤研究 (C)(18500073) により行なわれた．

[文献]

[1] R. Agrawal and R. Srikant: “Privacy-Preserving Data Mining.”, SIGMOD Conference, ACM, pp. 439–450 (2000).

[2] A. V. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke: “Privacy preserving mining of association rules.”, Inf. Syst., **29**, 4, pp. 343–364 (2004).

[3] R. Agrawal, R. Srikant and D. Thomas: “Privacy Preserving OLAP.”, SIGMOD Conference, ACM, pp. 251–262 (2005).

[4] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone: “Classification and Regression Trees.”, Wadsworth (1984).

[5] J. Han and M. Kamber: “Data Mining: Concepts and Techniques”, Morgan Kaufmann (2000).

[6] 福田剛志, 森本康彦, 徳山豪：“データサイエンスシリーズ 3 データマイニング”, 共立出版 (2001).

[7] 福田剛志, 森本康彦, 徳山豪：“多値属性を用いた最適なデータセグメンテーションを生成するアルゴリズム”, 電子情報通信学会技術研究報告, **98**, 316, pp. 83–91 (1998).

[8] Y. Morimoto, T. Fukuda, H. Matsuzawa, T. Tokuyama and K. Yoda: “Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases.”, VLDB, pp. 380–391 (1998).

高見澤 秀久 Hidehisa TAKAMIZAWA
群馬大学大学院工学研究科電子情報工学専攻博士後期課程在学中．2001 群馬大学工学部情報工学科卒．プライバシー保護等に興味を持つ．

有次 正義 Masayoshi ARITSUGI
熊本大学大学院自然科学研究科教授．1991 九州大学工学部情報工学科卒．1996 同大大学院了．博士 (工学)．同年群馬大学助手．同助教授を経て，2007 より現職．データベースシステム，分散並列データ処理等に興味を持つ．