プライバシー保護決定木学習におけるエントロピーを近似する順序同型関数

菊池浩明† 伊藤 孝一‡ 牛田 芽生恵‡ 津田 宏‡ 山岡裕司‡

†東海大学，東京都港区高輪 2-3-23, ‡富士通研究所，川崎市中原区上小田中 4-1-1

あらまし プライバシー保護決定木学習では，機密性のあるデータセットを持つ複数の組織が互いの値を秘匿した方法で，最適な識別子を選択するエントロピー利得を求める．しかし，この計算には大きな計算量を要する．そこで，この研究では，新たなエントロピー関数と順序同型な関数を最大値と最小値で定義し，それによる計算量削減を試みる．公開データセットにおける評価を報告する．

Order-Isomorphism Function for approximate entropy in Privacy-Preserving Decision Tree Learning

Hiroaki Kikuchi† Kouichi Ito‡ Mebae Ushida‡ Hiroshi Tsuda‡
Yuji Yamaoka‡

†Tokai University, 2-3-23 Takanawa, Minato, Tokyo, kikn@tokai.ac.jp
‡Fujitsu Laboratories Ltd., 4-1-1 Kamiodanaka, Nakahara, Kawasaki

Abstract Privacy-preserving decision tree learning protocol allow multiple parties with confidential datasets to jointly perform entropy gain to choose the best classifier in privacy-preserving way. The entropy function requires huge computational overhead to preform. Hence, in this study, a new order-isomorphism function is defined using simple max and min that are less intensive in computation. The evaluation with public datasets will be reported.

1 はじめに

互いに信頼していない A と B が，それぞれ条件属性 $a_1, \ldots, a_m$，と $b_1, \ldots, b_m$，目的属性 c_A と c_B についての，n 個の同期した垂直分割データセット (表 1) を持っているとする．互いのデータセットを秘匿したままで，論理演算で合成された目的属性，例えば，$c_A \wedge c_B$ や $c_A \vee c_B$ についての決定木を協力して求める問題を考える．

垂直分割データセットにおけるプライバシー保護の研究には，表 2 に示される代表的な研究がある．ここには大きく次の課題がある．

1. 目的属性をどのように管理するか
   両方が管理する [4] では，秘匿の度合いが弱く，協調作業で得られた木が同意できない時でもやり直しが出来ず，制約が多い．しかし， [2] の様に片側のみが管理するでは木の評価を行う際にも協調が必要になる．
2. 複雑で計算コストの高いエントロピー利得
   多くの決定木学習ではエントロピーの利得で識別に用いる属性を決めている．しかし，そのためには対数や実数を含む計算を必要とするので，制約のある秘密関数計算では現実的ではない．

そこで，本稿では，準同型性暗号を用いた内

表 1: 垂直分割データセット（A のデータセット D_A, B の D_B と目的関数 $f(c_A, c_B)$

a_1	a_2	c_A
1	1	1
1	0	1
0	1	1
1	1	0
0	1	0
0	0	0

b_1	b_2	c_B
1	1	0
0	1	1
0	1	1
1	1	1
1	0	0
0	0	0

$c_A \wedge c_B$	$c_A \vee c_B$	$\overline{c_A} \wedge c_B$
0	1	0
1	1	0
1	1	0
0	1	1
0	0	0
0	0	0

表 2: プライバシー保護決定木アルゴリズム

	Du and Zhan[4]	Vaidya and Clifton[2]	本提案
内積計算	TTP によるスカラー積プロトコル	準同型性暗号とセキュア内積プロトコル + SFE	セキュア内積プロトコル
識別の指標	Entropy or GINI 係数	Entropy	Entropy
目的属性	2 者間で共有	片側のみ	それぞれで保有
決定木の評価	公開，誰でも評価可能	秘密，2 者間協調で評価	2 者間協調
通信コスト	$4n$ 暗号なし	$n+1$	$n+1$

積プロトコルを基本として，分散管理された目的属性の論理積，論理和などの論理演算を許した新しい秘匿決定木アルゴリズムを提案する．エントロピーに代わる線形区分関数を導入することにより，エントロピーに基づく方式と同じ結果を効率的に計算できる特徴を持つ．

2 要素技術

2.1 決定木学習

属性の集合を $W = \{w_1, \ldots, w_m\}$, 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．

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．

2.2 論理決定木学習アルゴリズム ID3

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 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．

2.3 暗号要素技術

秘匿内積プロトコルをアルゴリズム 1 に示す．

2.4 秘密関数計算

任意の関数の秘匿計算プロトコル SFE を仮定する．

3 提案方式

3.1 情報量利得に代わる線形区分関数

エントロピーに基づく情報量利得は決定木学習に最も広く用いられている分割指標だが，その計算には対数を含む複雑な計算が欠かせない．確率が $p = x/y = u+v$ と 2 者間に分割されたときに，エントロピー　$H(S) = -\sum_i p_i \log(p_i)$ を求めることは，SFE に対数 $log((u+v)/n)$ の計算をさせなくてはならない．このコストを下げるために，Gini 係数 $Gini(S) = 1 - \sum_{p_i} p_i^2$ を導入したり [4], 対数のテーラー展開

$$\ln(1 + \epsilon) = \sum_i^k \frac{(-1)^{i-1} \epsilon^i}{i}$$

を用いたり [3] して工夫をしている．しかしながら，SFE はゲートレベルで暗号計算を繰り返すので，単純な加算や比較ならよいが，積や実数の計算は実用的ではない．そこで，情報量利得に代わる新たな指標を提案する．

定義 3.1 任意の $x, y \quad \in [0, 1]$ について，$g(x) > g(y)$ である時，その時に限り，$f(x) > f(x)$ である時，f は g と**順序同型** (order-isomorphism) であるという．

Algorithm 1 Secure Scalar Product

Input: Alice has n-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_n)$. Bob has n-dimensional vector $\boldsymbol{y} = (y_1, \ldots, y_n)$.

Output: Alice has s_A and Bob has s_B such that $s_A + s_B = \boldsymbol{x} \cdot \boldsymbol{y}$.

1. Alice generates a homomorphic public-key pair and sends the public key to Bob.
2. Alice sends to Bob n ciphertexts $E(x_1), \ldots, E(x_n)$.
3. Bob chooses s_B at random, computes

$$c = E(x_1)^{y_1} \cdots E(x_n)^{y_n} / E(s_B)$$

   and send c to Alice.
4. Alice decrypts c to get $s_A = D(c) = x_1 y_1 + \cdots + x_n y_n - s_B$.

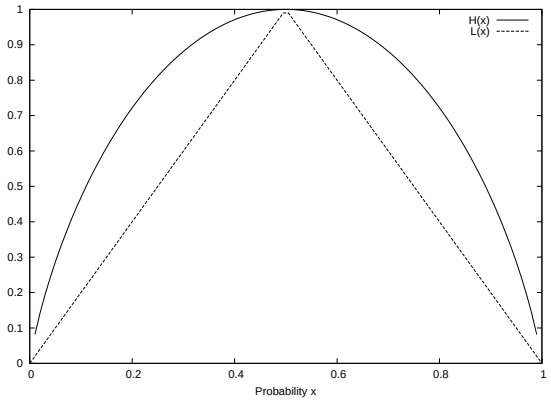

図 1: エントロピー関数 $\mathcal{H}(x)$ と線形区分関数 $L(x)$

2元情報源においては，生起確率が $p_1 = 1 - p_0$ なので，そのエントロピーは1変数のエントロピー関数 $\mathcal{H}(x) = -x \log x - (1-x)\log(1-x)$ で与えられる．$\mathcal{H}$ は，図1に示される $\mathcal{H}(0) = \mathcal{H}(1) = 0$, $\mathcal{H}(1/2) = 1$ となる凸連続関数である．そこで，この性質を保持した次の線形区分関数 $L:\ [0,1] \to [0,1]$ を考える．

$$L(x) = \begin{cases} 2x & \text{if } x < 1/2, \\ 2 - 2x & \text{if } x \geq 1/2. \end{cases} \tag{1}$$

命題 3.1 エントロピー関数 $\mathcal{H}(x)$ は，線形区分関数 $L(x)$ と順序同型である．

(証明) ある x, y について，$\mathcal{H}(y) < \mathcal{H}(x)$ であるとする．この時，(1) $y < x < 1/2$, (2) $1/2 < x < y$, (3) $y < 1 - x < 1/2 < x$, (4) $x < 1/2 < 1 - x < y$ のどれかである．(1) の時，$L(y) = 2y < 2x = L(x)$, (2) の時，$L(x) = 2 - 2x < 2 - 2y = L(y)$ であり，$L(y) < L(x)$ が成立する．(3), (4) の時も同様である． (証明終)

条件属性においてデータセットを条件付けるときには，条件属性の値に応じたエントロピー関数の期待値

$$H(S|a) = \frac{\boldsymbol{a}^2}{n} H(S|a=0) + \frac{n - \boldsymbol{a}^2}{n} H(S|b=0)$$

を求める必要がある．従って，L が決定木学習の観点でエントロピーと同値であることを示すには，$\mathcal{H}$ と L が順序同型であるだけでは十分ではない．

命題 3.2 a を条件属性，$\boldsymbol{a}^2$ をその属性を満たす行数，$P(c|a)$ を a で条件付けられたある属性 c の確率とする．$\boldsymbol{a}^2 L(P(c|a))$ は整数である．

(証明) $P(c|a) = p = x/y = x/\boldsymbol{a}^2$ と置くと，$p < 1/2$ の時，$\boldsymbol{a}^2 L(p) = y2(x/y) = 2x$. $p \geq 1/2$ の時も同様. (証明終)

よって，これを，

$$L_2(x/y, y) = \begin{cases} 2x & \text{if } 2x < y, \\ 2y - 2x & \text{if } 2x \geq y \end{cases} \tag{2}$$

と置く．この時，次の性質が成立する．

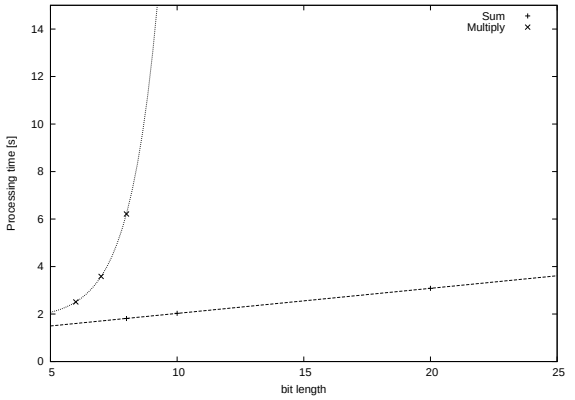

図 2: 秘密関数計算 fairplay における加算比較と乗算比較の処理時間

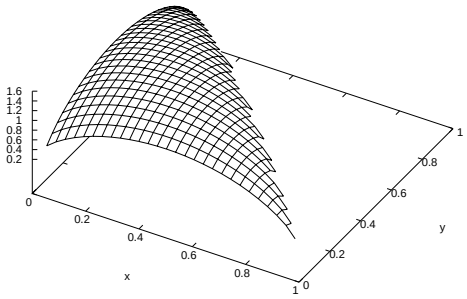

図 3: 3 値の目的属性におけるエントロピー

命題 3.3 データセット D において，条件属性 a, b と目的属性 c が 2 値である．$H(c|a) > H(c|b)$ である時，かつその時に限り，$L_2(P(c|a=1), \boldsymbol{a}^2) + L_2(P(c|a=0), \overline{\boldsymbol{a}}^2) > L_2(P(c|b=1), \boldsymbol{b}^2) + L_2(P(c|b=0), \overline{\boldsymbol{b}}^2)$.

こうして，2 値のデータセットに限っては，エントロピー関数 $\mathcal{H}$ で計算した決定木と線形区分関数 L で計算した決定木は同一であることが示された．

線形区分関数は，期待値の計算過程で L_2 が整数となり，最適属性を求めるのも，整数の和の比較プロトコルでよいので SFE で十分に計算可能である．例えば，代表的な SFE の実装処理系である fairplay[7] における処理時間は図 2 によると，20 bit の加算の比較 $(u+v > u'+v')$ は 3 秒以内で実行できる．

3.2 3 値以上のデータセットにおける線形区分近似

しかしながら，この性質が保証されるのは目的属性の値がブール値などの 2 値あることを仮定している．UCI Machine Learning Repository に格納されている 28 種類の識別データセットの中で，目的属性が 2 値のものには，Baloon (T/F), Breast Cancer, Chess, Congressional Voting (民主党, 共和党), LED Display, MONK, Mushroom (食用，毒性), Shuttle (自動，手動), SPECT Heart, Tic-Tac-Toe (正，負), Train(東

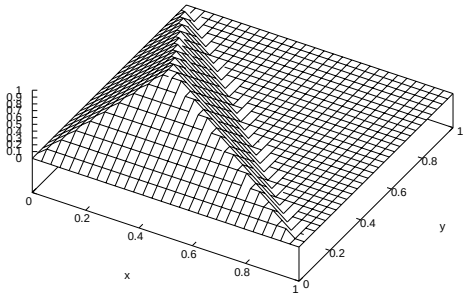

図 4: エントロピーを線形近似する提案テスト指標 $L(x)$

行，西行) などがあり，多くはこの手法がそのまま適用できる．

3 値以上ある目的属性においても，線形区分関数を定義できる．しかし，こちらはエントロピーに基づく比較と線形区分関数の比較が完全に一致する保証はない．3 値のエントロピー関数 $\mathcal{H}_3(x, y, 1-x-y)$ の分布を図 3 に示す．$x = y = z = 1/3$ の時，最大値 $\mathcal{H}(1/3, 1/3) = \log_2(3)$ を取り，x, y, z のいずれにおいても，$\mathcal{H}_3(0, \cdot, \cdot) = \mathcal{H}_3(1, \cdot, \cdot) = 0$ である．これを近似する線形区分関数には，例えば，

$$L_3(x,y,z) = \begin{cases} 3y & \text{if } y < x,\ 1 \geq x + 2y, \\ 3x & \text{if } y \geq x,\ 1 \geq x + 2y, \\ -3x - 3y + 3 \vee 0 & \text{otherwise.} \end{cases} \tag{3}$$

で定義できる図 4 に示す関数が定義できる．

3.3 提案方式

(Input) A は，目的属性 c_A を有するデータベース D_A を持ち，同様に B は D_B を持つ．f は目的属性の論理式とする．A (B) の管理する最適属性の集合 $E_A = E_B = \emptyset$ と初期化する．

1. A は目的とするターゲット属性の関数 $f(c_A, c_B)$ の，条件属性 a_i $(i = 1, \ldots, m_A)$ についての条件確率 $P(f(c_A, c_B) = 1 | a_i, E_A)$ を式 (**??**), (**??**) で与えられる B との秘匿内積プロトコルで $(u_i + v_i)/(\boldsymbol{a_i}^2) = P(f(c_A, c_B) = 1 | a_i, E_A)$ となる u_i と $\boldsymbol{a_i}^2$ を求める．この時 B は v_i を得る．同様に，B の持つ条件属性 b_i $(i = 1, \ldots, m_B)$ についても秘匿内積プロトコルを実行して，v'_i, $\boldsymbol{b_i}^2$ を得る．A も u'_i を得る．

2. A, B は, 秘密関数計算プロトコル FSE を用いて，$u_1, \ldots, u_{m_A}, u'_1, \ldots, u'_{m_B}, \boldsymbol{a_1}^2, \ldots, \boldsymbol{a_{m_A}}^2$, $u_1, \ldots, u_{m_A}, u'_1, \ldots, u'_{m_B}, \boldsymbol{b_1}^2, \ldots, \boldsymbol{b_{m_B}}^2$ を入力し，全ての条件属性 a_i について，$L_2(u_i + v_i, \boldsymbol{a_i})$ を計算し，情報量利得を最大化する最適属性 a_* (B の条件属性の時は，b_*) を出力する．

3. 最適属性の集合 $E_A = E_A \cup \{a_*\}$ を更新し，終了条件[1]を判定し，条件を満たさない時は，1 から繰り返す．

(Output) 最適属性の集合 E_A, E_B.

4 おわりに

区分線形関数で定義された識別基準について，安全に，かつ高速に識別に用いる条件属性を決定することが出来る新しい垂直分割データセットにおける決定木学習アルゴリズムを提案した．

[1] その条件のクラスがすべて同一である時，枝刈りによって指定される識別の最少事例数を下回った時，識別に用いる条件属性が存在しない時などで決まる

参考文献

[1] G. Jagannathan, K. Pillaipakkamnatt, R. N. Wright, A Practical Differentially Private Random Decision Tree Classifier (ICDM '09)

[2] J. Vaidya and C. Clifton, Privacy-preserving decision trees over vertically partitioned data, 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, 2005.

[3] Lindell, Y., Pinkas, B., Privacy Preserving Data Mining Advances in Cryptology - CRYPTO 2000 Lecture Notes in Computer Science 1880, Springer, pp. 36-54, 2000.

[4] W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data", IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, Vol. 14, 2002.

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

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

[7] Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella, "Fairplay – A Secure Two-Party Computation System", Usenix Security Symposium, 2004.