情報処理学会論文誌　データベース　Vol. 4　No. 2　126–141 (July 2011)

プライバシ保護データ公開に向けた l-多様化適性の評価

村本俊祐[†1,*1]　上土井陽子[†1]　若林真一[†1]

本稿ではプライバシ保護データ公開のために利用される一般化に基づくデータ変換問題に着目する．プライバシ保護データ公開のためのデータ変換として k-匿名化が一般的であった．しかしながら，近年 k-匿名化データの脆弱性が指摘されており，より高い安全性を保証するためのデータ変換として l-多様化が注目されている．l-多様化では k-匿名化と比較してデータが一般化される度合が高くなり，データ損失が大きくなることが問題である．我々が行った予備実験では発見的手法を適用して l-多様化したときのデータ損失は k-匿名化したときのデータ損失と比較して数倍ほど大きかった．匿名化と比べ多様化のデータ損失が大きい原因を明らかにするため入力データの l-多様化に関する性質の解析が必要となる．本稿では変換対象のデータを k-匿名化した場合を基準として l-多様化に適しているかを判定するために用いる 3 つの指標を提案し，それら指標の有効性について考察することを目的とする．

Evaluation of Property of l-diversification toward Privacy-Preserving Data Publishing

Shunsuke Muramoto,[†1,*1] Yoko Kamidoi[†1]
and Shin'ichi Wakabayashi[†1]

In this paper, we focus on an privacy protection technique that uses the generalization of data on an input data table. One of the privacy protection techniques is k-anonymization. However, attacks on k-anonymitized data are known and a novel privacy criterion called l-diversity is alternatively focused in recent years. In our experimental simulations, our proposed l-diversity heuristic algorithm found solutions with about 2–10 times worse information loss than ones by our k-anonymity heuristic algorithm. In order to investigate some factor, we need metrics to evaluate applicability of l-diversification of data. In this paper, we derive a lower bound of an optimal l-diversification cost. Next, we induce the maximum number of blocks and a lower bound of maximum block sizes of solutions of l-diversification of data. We try to evaluate adequacy of data for l-diversification by using proposed metrics.

1. はじめに

近年，プライバシを保護したうえで官庁，企業，医療機関等が蓄積しているデータを公開し，様々な角度からの解析を可能にしようとするプライバシ保護データ公開（Privacy-Preserving Data Publishing: PPDP）が注目されている[2]．プライバシ保護データ公開ではデータの個人情報を保護したうえでデータを公開するために，データ解析において元データを入力としたときと同程度の有益な情報を得られるようデータを変換する問題が考察される．プライバシ保護データ変換問題に対する基本方法として，暗号を利用する方法やデータが k-匿名性という性質を保持するようデータを一般化する k-匿名化手法が広く知られている．しかしながら，k-匿名性保持データに対する新しい攻撃タイプの存在が指摘され，さらなる保護強化のため l-多様性という性質保持を公開データに課する l-多様化手法が注目を集めている[7),8)]．

k-匿名化手法や l-多様化手法はデータ一般化というある値をより一般的な値に換えるデータ変換により，データを復元不可能にしてプライバシを保護する．暗号を利用する方法とは異なり，データ一般化を用いる場合，データを受け取る相手，つまり，データを解析する機関内の関係者が攻撃者になる可能性を想定している[2]．一方，暗号を利用するプライバシ保護方法では解析者が攻撃者になることを想定していない．暗号化を利用したデータ変換方法と異なりデータ一般化による方法では元のデータが持っている有用性をできるだけ保ったままでプライバシ保護要求を満たすことが目的となる．元データの有用性を損なう可能性のある情報損失をデータが歪曲された度合（データ歪曲度）として表現し，データ変換における最小化目的関数として定義する手法が多く存在する[1),14)]．データ歪曲度の最小化を目的とする k-匿名化問題は NP 困難であり，これより l-多様化問題も NP 困難といえる[1)]．l-多様性に関する理論的解析結果として文献 15) がある．文献 15) では文献 1) で扱った NP 困難性証明の対象となった k-匿名化問題の自然な拡張としての l-多様化問題よりさらに限定した問題でも NP 困難であることを示している．また，同文献では l-多様化問題に対する近似アルゴリズムを提案しているが，情報の段階的な一般化ではなくまったく値を表示しない抑制を用いてデータを多様化するため評価関数として一般に扱われている情報損失の

†1 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University
*1 現在，日本電気株式会社
Presently with NEC Corporation

コストを単純化した関数を用いている．文献 5) では l-多様化の限界として注目属性値の分布が非常に偏っている入力テーブルで l-多様化を達成することが難しいことを例によって説明している．しかしながら，入力テーブルによる l-多様化の困難さの違い等を明らかにする定性的な解析はなされていなかった．一般に l-多様化は k-匿名化に比べ高いプライバシ保護要求を満たすことができるが，より大きなデータ歪曲を必要とする．我々の行った予備実験ではデータによっては l-多様化した解のデータ歪曲度は k-匿名化の解のデータ歪曲度の 10 倍以上になる場合もあり[11]，定性的な解析が求められる．

本稿では，入力データの k-匿名化と比較した l-多様化適性の評価に用いる指標として，l-多様化問題における最適歪曲度の下界，ブロック数の最大値，最大ブロックサイズの下界という指標の算出方法を提案し，データに対して提案指標を用いて l-多様化適性を評価する．

本稿の以降の構成は次のとおりである．2 章で準備として用語の定義，k-匿名性，l-多様性について定義する．3 章では l-多様化適性を評価するための指標を提案し，4 章において提案指標の有効性を検証する．最後に 5 章において適性評価を取り入れたプライバシ保護システム案を応用例として示す．

2. 準　　備

2.1 データテーブル

本稿ではデータテーブルとして**表 1** のような有限個のタプル（行に対応）と属性（列に対応）からなるものを考慮する．ここで各タプルは各属性に属するデータ値の $n+1$ 個の組とする．$n+1$ を属性数と呼ぶ．このうち，データ解析者が注目していて一般化を行ってほしくない属性を注目属性（Sensitive Attribute）という．その他の属性を非注目属性（Non-Sensitive Attribute）という．たとえば，表 1 では注目属性は Condition であり，その他の属性は非注目属性である．非注目属性のうち，データ推測から秘密にしたい情報（ここでは個人）を特定する可能性のある，単独の識別子ではないが組み合わせることで同じ働きをする恐れのある属性の集合を準識別子 QI と呼ぶ．たとえば，表 1 では QI= {Race, Birth, Gender, ZIP} である．k-匿名性，l-多様性を保持することを目的に行うデータ変換では個人を識別できないようにするため準識別子の属性の値を一般化する．注目属性が多いほど，データが持つ注目属性値の組合せの多様さは大きくなり一般に l-多様化問題は簡単になる．よって，本稿では注目属性は 1 属性と仮定する．以降では注目属性を除いたデータテーブルの属性数を n で表すとする．

2.2 k-匿名性[14]

本稿ではデータテーブルの k-匿名性を以下のように定義する．

データテーブル中の各タプルにおいて，そのタプルの持つ準識別子 QI のデータ値組合せ（各属性値の組合せ）と同じ準識別子 QI のデータ値組合せを持つタプルが自分自身を含め k 個以上存在する状態

k-匿名性を保持するテーブルの例をあげる．表 1 (a) のテーブル PT が与えられたとき，表 1 (b) のテーブル RT に変換したとする．テーブル RT では，$t1'$，$t2'$ のタプルが同一データ値組合せを持っており，同様に $t3'$，$t5'$，$t7'$ の 3 つのタプル，$t4'$ と $t6'$ の 2 つのタプルがそれぞれ準識別子 QI において同一データ値組合せを持っている．よって，テーブル RT ではすべてのタプルで準識別子 QI に関し同一データ値組合せを持っているタプル

表 1　2-匿名性保持を目的とした一般化
Table 1　An example of 2-anonymization by generalization.

(a) 初期テーブル PT

	Race	Birth	Gender	ZIP	Condition
t1	African	1964	female	02138	H.D.
t2	African	1964	female	02138	Cancer
t3	African	1967	male	02141	V.I.
t4	European	1971	female	02139	Cancer
t5	European	1967	male	02141	H.D.
t6	European	1971	male	02139	V.I.
t7	European	1965	male	02141	V.I.

(b) 2-匿名化テーブル RT

	Race	Birth	Gender	ZIP	Condition
t1'	African	1964	female	02138	H.D.
t2'	African	1964	female	02138	Cancer
t3'	Person	196*	male	02141	V.I.
t4'	European	1971	human	02139	Cancer
t5'	Person	196*	male	02141	H.D.
t6'	European	1971	human	02139	V.I.
t7'	Person	196*	male	02141	V.I.

V.I.=Viral Infection, H.D.=Heart Disease

が自分を含め 2 個以上存在する．このとき，テーブル RT は 2-匿名性を保持しているという．また，本稿では，k-匿名性を保持させるために行うテーブル変換を k-匿名化と呼ぶ．

2.3 k-匿名化では防げない攻撃

データ値組合せによるデータ推測は k-匿名化することで防止できると考えられていた．しかし，近年では k-匿名化によっても防ぎきれない攻撃が存在することが指摘されている．そのような攻撃として以下の 2 種類の攻撃があげられている．

- 同種攻撃（Homogeneity Attack）
- 背景知識攻撃（Background Knowledge Attack）

実際に文献 7) において，注目属性に出現する値の種類が 3 種類，タプル数 60000 のデータテーブルにおいて 5-匿名性を保持させた場合，最低 740 人以上の個人データが新種の攻撃により推測される危険性があるとされている．上記の攻撃では匿名化後の準識別子 QI において同一データ値組合せを持っているタプルの注目属性値に着目している．同種攻撃，背景知識攻撃とも攻撃者は対象者についての準識別子の属性値と公開された表中に対象者のタプルが含まれていることを知っているとする．

同種攻撃は匿名化後の表のタプルグループの注目属性値が同種である場合に対象者の注目属性値を一意に特定する攻撃である．例として，**表 2** (a) のテーブルが与えられ，表 2 (b) の 4-匿名を持つテーブルに変換されたとする．確かに表 2 (b) の結果では，Bob のデータがどのタプルかという確証は得られない．しかし，先の条件より攻撃者は Bob がどこのタプルグループに含まれているか推測することが可能である．つまり，表 2 (b) の結果において攻撃者は Bob のデータが含まれるであろうと推測されるタプルグループの注目属性がすべて Cancer であることから，Bob は Cancer を患っていると特定できる．

背景知識攻撃は同種攻撃ほど一意に対象者の注目属性値を特定できないが，対象者についての背景知識を利用して注目属性値の候補を絞り込む．仮に “日本人は Heart Disease の発病率が非常に低い” という背景知識を持っているとする．攻撃者が Umeko の非注目属性の値を知っていて，さらに “日本人である Umeko” のデータは注目属性の値が Heart Disease であるタプルには含まれない確率が高いことから表 2 (b) 中の上から 1，2 番目のタプルが Umeko の候補から外れる．したがって，高い確率で Umeko の病名は Viral Infection であると推測されてしまう．

2.4 l-多様性[7),8)]

非注目属性の属性値において，一般化されたデータが準識別子 QI において同一データ値組合せを持ち，その組合せが仮に $q*$ となる場合のタプルグループをまとめて $q*$-ブロック

表 2　新種攻撃の例
Table 2　An example of the new attacks.

(a) 初期テーブル

ZIP	Nationality	Condition	Name
13053	Russian	Heart Disease	
13053	American	Heart Disease	
13052	Japanese	Viral Infection	Umeko
13052	American	Viral Infection	
13065	American	Cancer	Bob
13068	India	Cancer	
13067	Japanese	Cancer	
13068	American	Cancer	

(b) 4-匿名化テーブル

ZIP	Nationality	Condition	Name
1305*	*	Heart Disease	
1305*	*	Heart Disease	
1305*	*	Viral Infection	Umeko
1305*	*	Viral Infection	
1306*	*	Cancer	Bob
1306*	*	Cancer	
1306*	*	Cancer	
1306*	*	Cancer	

と呼ぶことにする．k-匿名性を保持しただけでは新種の攻撃を防ぐことができなかった理由として次の 2 点があげられる．

- $q*$-ブロック中の注目属性の多様性の欠如
- 強力な背景知識

これらの原因による攻撃への対処のために導入された l-多様性の定義は次のように表すことができる．ここでの l の多様性という概念については後に詳しく定義する．

テーブル中のすべての $q*$-ブロックにおいて出現する注目属性の値が少なくとも l の多様性を持つ状態を保つ

本稿では，l-多様性を保持させるために行うテーブル変換を l-多様化と呼ぶ．また，l-多様化されたテーブルでの各ブロックのサイズは l 以上であることは自明であり，l-多様化されたデータテーブルは，同時に l-匿名化されている．l の多様性を保持している状態を詳細に定義するのに次の方法があげられる[7),8)]．

- 単純 l-多様性
- エントロピー l-多様性
- 再帰的 (c, l)-多様性

初めの単純 l-多様性ではブロックの多様さを単にブロックに含まれる注目属性値の種類数によって表すとする．エントロピー l-多様性の定義によるプライバシ保護要求は単純 l-多様性，再帰的 (c, l)-多様性の定義によるプライバシ保護要求よりもより厳しいといわれている[8)]．文献 11) で我々が提案した l-多様化アルゴリズム *DiverDIS* において，l-多様性の定義としてエントロピー l-多様性と再帰的 (c, l)-多様性を使用した場合のシミュレーション実験を行ったが，データ歪曲度に対して大きな差はなかった．また，再帰的 (c, l)-多様性は入

力定数 c により性質も変動するため，本稿では，定性的な評価の対象としてエントロピー l-多様性について注目し，以下にその定義を示す．

2.4.1　エントロピー l-多様性

エントロピー l-多様性ではエントロピーを用いてデータの多様さを表す．データテーブルにおける l-多様化では注目属性中に出現する各種類の値を各ブロックに満遍なく散らばせることでエントロピーを最大に近づけることが可能となる．

l-多様化後のテーブル PT は準識別子 QI において同じデータ値組合せ $q*$ を持つタプルの集合に分割できる．テーブル PT の $q*$-ブロックへの分割をブロックの集合 $B = \{b|\ b$ は $q*$-ブロック $\}$ とする．ここで $\bigcup_{b \in B} b = PT$，かつ，B に含まれる任意の異なるブロックのペア b，b' において $b \cap b' = \emptyset$ とする．以下の式を満たす最大の l の値が $q*$-ブロックの保持している多様さ l の値となる．以下ではブロック分割 B の大きさ $|B|$ をブロック数，各ブロック $b \in B$ の大きさ $|b|$ をブロックサイズと呼ぶ．式中の関数 $P(q*, s)$ は $q*$-ブロック中において準識別子に対応する属性の値の組合せが $q*$ かつ注目属性の値が s であるタプルの出現頻度（割合）を表すものとする．また，注目属性の値の母集合を S とする．

$$-\sum_{s \in S} P(q*, s) \log(P(q*, s)) \geq \log(l)$$

テーブル中のすべての $q*$-ブロックについて上記の式が満たされるとき，テーブルはエントロピー l-多様性を保持しているという．つまり，l-多様化後のテーブル中のすべてのブロックのエントロピーに関する多様さの中での最小値がブロック分割 B の多様さとなる．

ここで，注意すべき点として，変換テーブルがエントロピー l-多様性を満たしているとき各ブロックは少なくとも l 種類の注目属性値を含むが，l 種類の注目属性値がブロック中に存在していても必ずしもエントロピー l-多様性を保持しているとはいえないことがあげられる．たとえば，ブロック中のある注目属性値の出現頻度が他の注目属性値に比べて極端に高い，または極端に低い場合，注目属性値の種類数以下の多様性となる．さらに，ブロック中の注目属性値の種類数より大きなエントロピー多様性をブロックが保持することはない[8]．ゆえに，エントロピー l-多様性を保持させるには，なるべくブロック中の l 種類以上の注目属性値の出現頻度を均等にすることが望ましいとされる．

2.5　データ歪曲度算出関数 *DIS*

データテーブルを k-匿名化または l-多様化させるためには，一般化等のデータ操作を必要とし元のデータを歪曲してしまう．そこで，テーブル変換の解析への影響を小さくするため，元のテーブルになるべく近い形でテーブルを変換することが k-匿名化，l-多様化の目標

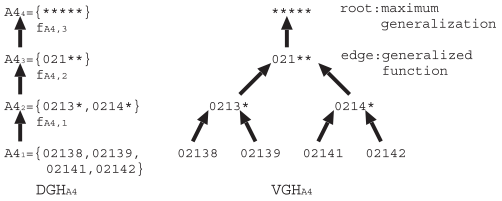


図 1　属性 A4 (ZIP) の属性一般化階層 DGH，値一般化階層 VGH
Fig. 1　DGH and VGH for Domain A4 (ZIP).

となる．よって，元データテーブルからどの程度，情報が失われたか（データ歪曲度）を評価するため，文献 14) をもとに，より形式的に再定義したデータ歪曲度算出関数 *DIS* をデータ変換の評価に用いる[9),10)]．

データ歪曲度算出関数 *DIS* では，属性を初期値の集合から最大一般化値までに一般化された回数で階層的に分ける属性一般化階層 *DGH*（Domain Generalization Hierarchies）と，一般化前の値と後の値の関係を木（最大一般化値を根とする）で表現した値一般化階層 *VGH*（Value Generalization Hierarchies）という一般化表現を使用する．表 1 のデータテーブル中の属性 *ZIP* に関しての属性一般化階層 *DGH* および値一般化階層 *VGH* の例を**図 1** に示す．属性一般化階層 *DGH*，値一般化階層 *VGH* はデータテーブル上の各属性それぞれに対して定義され，複数階層からなる．基本的に属性一般化階層 *DGH* と値一般化階層 *VGH* はデータテーブルの管理者が作成するものとする．また DGH_{Ai}，VGH_{Ai} は属性 Ai の属性一般化階層 *DGH* と値一般化階層 *VGH* とする．このとき，テーブル PT の j 番目のタプル tj の属性 Ai の値が一般化によりテーブル RT の j 番目のタプル tj' の属性 Ai の値に変換されたときのデータ歪曲コスト $CA(Ai, tj, tj')$ は次のようになる．

$$CA(Ai, tj, tj') = \frac{h(VGH_{Ai}, tj(Ai)) - h(VGH_{Ai}, tj'(Ai))}{|DGH_{Ai}|}$$

ここで *DGH* の絶対値は属性一般化階層関数 *DGH* の階層数を表すとする．式中の tj はタプルを示し，$tj(Ai)$ でタプル tj 中の属性 Ai に対応する値を示し，関数 $h(tree, v)$ は木 $tree$ 中の値 v の高さを返す関数とする．

データ歪曲コストを用いてテーブル PT が一般化テーブル RT に変換されたときのデータ歪曲度算出関数 *DIS* の定義式を以下に示す．ここでも，テーブル PT の j 番目のタプル tj はテーブル RT の j 番目のタプル tj' に対応しているとする．この定義式ではテーブル全体の歪曲コスト（歪曲度の総和）を比較のため要素数で割り，正規化を行っている．

$$DIS(PT, RT) = \frac{\sum_{Ai \in QI} \sum_{tj \in PT} CA(Ai, tj, tj')}{|PT| \cdot |QI|}$$

一般化テーブル RT が一般化される前のテーブル PT とデータ値がまったく同じであれば $DIS(PT, RT)$ は 0 となる．また，一般化が行われるにつれて数値は大きくなり，すべてのデータ値が完全に抑制された状態（すべてが * になる等，情報が得られない状態）だと $DIS(PT, RT)$ は 1 となる．したがって，データ歪曲度算出関数 DIS は 0 から 1 の値を取る．属性一般化階層，値一般化階層を効率良く決定するための手法[3),4)] が存在する．本稿ではこれらの階層は入力として与えられるものとする．

2.6 k-匿名化問題と l-多様化問題

前節までに述べたように，プライバシ保護データを外部の解析者等に公開する場合，解析者の立場からは k-匿名性または l-多様性を満たすよう最小のデータ歪曲度でデータ変換が望ましい．

本稿では，入力テーブル PT を k-匿名性，l-多様性を保持するような結果テーブル RT をデータ歪曲度 $DIS(PT, RT)$ の最小化を目的として求める問題をそれぞれ k-匿名化問題，l-多様化問題と呼ぶ．このとき，各プライバシ保護条件を満たすデータ歪曲度が最小の結果テーブルを最適解と定義する．また，最適解の歪曲度にテーブル PT のタプル数（テーブルサイズ）を乗じた値を最適歪曲コストと呼ぶこととする．入力テーブルに対する一般化テーブルが k-匿名性，l-多様性の保持という制約条件を満たすとき，そのテーブルを許容解と定義する．許容解のうち，k-匿名性，l-多様性の保持において無駄なデータ変換を含まない解を極小解とする．

3. l-多様化適性の評価

我々は以前に，データ歪曲度算出関数 DIS を取り入れた発見的手法である k-匿名化アルゴリズム $MinDIS$[9)] と l-多様化アルゴリズム $DiverDIS$[11)]（ともに極小解を導出）を提案した．それらの提案手法を用いて予備実験を行い，それぞれの導出解のデータ歪曲度を比較した[11)]．シミュレーション実験の結果より，我々は入力データが l-多様化に適したデータであるかないか，適性を評価する必要があると考えた．そこで，我々は適性評価を行う判断指標として，k-匿名化問題に対する文献 1) の結果を拡張し単純 l-多様化問題における最適歪曲コストの下界，単純 l-多様化問題における許容解が持つブロック数の最大値，エントロピー l-多様化問題における許容解の最大ブロックサイズの下界の 3 つの指標値を求める方法を提案する．

3.1 l-多様化問題の最適歪曲コストの下界

l-多様化問題をグラフ問題に帰着させて最適歪曲コストの下界を導出する．テーブル PT を重み付き有向完全グラフ G=(V, E) に変換する．テーブル PT 中の各タプルはグラフ G 上の 1 つのノードに 1 対 1 に対応し，ノード間の枝には歪曲コストに関する重みを持つ．また，グラフ G 上の部分木によって，同一データ値組合せを持つブロックを表現することができる．以下に示すコストと近傍の定義を用いて，最適歪曲コスト OPT の下界を導出する．

3.1.1 コストの定義

定義 1　枝コスト

ノード間の枝のコストは，枝の始点ノードと終点ノードに対応する 2 つのタプルを同一データ値組合せを保持するデータに変換した場合に生じる始点ノードのデータ歪曲度と定義する．枝 (u, v) の枝コストを $CE(u, v)$ と表す．また，$CE(e)$ により枝 e のコストを示す．

定義 2　ノードコスト

l-多様化によって構成された同一データ値組合せとするブロックを表現するグラフ上の木において，木（ブロック）に含まれるすべてのノードの非注目属性値の組合せを最小限のデータ歪曲で同一にした場合に生じる各ノードのデータ歪曲度をノードコストと定義する．ノード u のコストを $CV(u)$ と表す．ノードコストはブロックを同一データ値組合せにするためのデータ歪曲度であることよりブロック中のノード u のコストと木に含まれる任意の他のノード v との間の枝コスト $CE(u, v)$ には以下の関係が成り立つ．

$$CV(u) \geq CE(u, v) \tag{1}$$

3.1.2 近傍の定義

l-多様化の適性評価指標の導出において，重要な点として，l-多様化では k-匿名化と違い，単純に近いノードどうしを接続してブロック（木）を作成しても多様性の向上は保障されないという点があげられる．よって，l-多様化の場合，注目しているノードに対する近傍をノードを対象として探すより，属性値を対象として決定しなければならない．本稿では属性値を対象として近傍を以下のように定義する．

定義 3　近傍属性値

あるノード u に注目したとき，注目属性値 $s \in S$ を持つノードの中でノード u に最も近い（枝コストが小さい）ノードから u への距離を $CES(u, s)$ で表すとする．つまり，$CES(u, s) = \min\{CE(u, v) | v \in V - \{u\} \wedge v$ の注目属性値 $= s\}$ である．このとき，ノード u の注目属性値 s_u と異なる注目属性値の集合 $S - \{s_u\}$ に属するすべての値 s を距離

$CES(u,s)$ を key として非減少順にソートしたとき，k $(1 \leq k \leq |S|-1)$ 番目にある注目属性値 s_k をノード u の k-近傍属性値といい，$s(u,k)$ で表す．

3.1.3　最適歪曲コスト OPT の下界

補題 1　入力データテーブルの l-多様化の最適歪曲コストを OPT と定義し，最適解におけるノード u のノードコストを $CV_{OPT}(u)$ とする．l-多様化の最適歪曲コスト OPT において，次の関係が成り立つ．

$$OPT = \sum_{u \in V} CV_{OPT}(u) \geq \sum_{u \in V} CES(u, s(u, l-1))$$

ここで，$s(u,l-1)$ はノード u の $(l-1)$-近傍属性値を表すとする．

[証明]　データテーブル PT が与えられた場合の l-多様化の最適解の下界について考える．テーブル PT 中のタプルに対応するノード u に注目したとき，最適解においてノード u に対応するタプルが含まれるブロックを $B_{opt}(u)$ とするとそのブロックはノード u を根として，ブロックに含まれるノード u 以外のノードを u の直接の子とするような木として表現できる．このとき，$v \in B_{opt}(u) - \{u\}$ において，式 (1) より $CV_{opt}(u) \geq CE(u,v)$ がいえる．l-多様性を満たすにはブロック $B_{opt}(u)$ に属するノードは少なくとも l 種類の異なる属性値を持たなければならない．よって，ノード u に l-多様性を保持させる場合に発生するデータ歪曲コストは，ノードの組合せに基づき，どのような同一データ組合せブロックを作成した場合でも，ブロックを表現する木において少なくとも 1 本の枝はノード u の $(l-1)$ 近傍属性値 $s(u,l-1)$ を持つノードとの間の最小の枝コスト $CES(u,s(l-1))$ 以上となる．したがって，最適歪曲コストは各ノード u とその $(l-1)$ 近傍属性値を持つノードとの最小距離 $CES(u,s(u,l-1))$ の総和より小さくなることはない．よって補題 1 が成り立つ．□

3.2　ブロック数の上界

l-多様化を行う場合，同一データ値組合せとして表現されるブロックには少なくとも l 個のタプルが含まれなければならない．しかし，注目属性値の分布が不均一な入力テーブルを l-多様化する場合には，最も大きなブロックには l 個よりもずっと多くの数のタプルが含まれる必要があるかもしれない．ある入力テーブルに対する l-多様化問題の許容解が持つブロックの数のうち，最大のブロック数を最大ブロック数と呼ぶとすると最大ブロック数も入力テーブルの注目属性値の分布に依存する．

l-多様化の許容解において構成されるブロックの最大サイズの下界は，解のブロックの数が多いほど小さくできる．よって，入力データテーブルに対して，ブロックの数を最大にできれば構成されるブロックの最大サイズの下界を求めることができる．そこで，単純 l-多様化問題の許容解が持つブロック数の最大値 $MaxNB$ を導出する．エントロピー l-多様化問題では $MaxNB$ は許容解が持つブロック数の上界である．また，入力テーブルにおける総タプル数 N を次の補題 2 で得られたブロック数の上界 $MaxNB$ で割った値の切り上げ値 $\lceil N/MaxNB \rceil$ が，最大ブロックサイズの下界 α となる．このとき，許容解では $\lceil N/MaxNB \rceil = \alpha$ 未満の最大ブロックサイズ，$(\alpha-1)$ 未満の平均ブロックサイズを持つことがないといえる．つまり，入力テーブルによっては k-匿名化問題と l-匿名化問題の解のデータ歪曲コストを比較する場合に $k=l$ として比較することは厳しすぎ，$k=\alpha$ としたときの α-匿名化問題の解との比較がより現実的であると予測される．

そこで，入力テーブルに対し，ブロック数の上界 $MaxNB$ を求め，l-多様化する場合に許容解が持つ最大ブロックサイズの下界 α を求める．以降の補題，定理で共通に用いる入力データテーブル PT の定義をまず示す．

定義 4　入力データテーブル PT

入力データテーブル PT を注目属性値として P 種類を持つテーブルとし，注目属性値の各種類ごとのタプル数の非増加順に種類を番号づけ $0, 1, \ldots, P-1$ とし，各種類 i のタプル数を N_i（$N_0 \geq N_1 \geq \ldots \geq N_{P-1}$）とする．また，

$$S_i = \sum_{i \leq j < P} Nj$$

と S_i（S_0 はタプル総数に対応）を定義する．

補題 2　注目属性値を P 種類持つ入力データテーブル PT，$l \leq P$ なる自然数 l において，

$$I = \min\{i | \lfloor \frac{S_i}{l-i} \rfloor \geq N_i, 0 \leq i < l\}$$

と定義した場合，入力データテーブル PT に対するブロック数が $MaxNB = \lfloor \frac{S_I}{l-I} \rfloor$ より大きな単純 l-多様化問題，および，エントロピー l-多様化問題の許容解は存在しない，かつ，ブロック数が $MaxNB$ の単純 l-多様化問題の許容解が存在する．

[証明]　単純 l-多様さの値 l に着目した帰納法により証明する．

(基本ステップ)

$l=1$ としたときに命題が成り立つことを示す．入力データテーブルの注目属性値の種類数 P が $P \geq 1$ を満たすことより，単純 1-多様性を保持させる場合は，単独のタプルでブロックを構成させることで最大ブロック数を達成できることから，最大ブロック数が S_0 であることは自明である．一方，$S_0/(1-0) \geq N_0$ となり，$I=0$ より $\lfloor \frac{S_0}{l-0} \rfloor = S_0$ より命題が成り立つ．

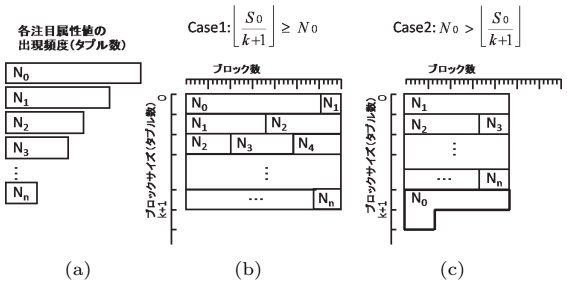


図 2　最大ブロック数の例
Fig. 2　Examples of solutions having the maximum number of blocks.

（帰納的ステップ）

$l = k$ としたときに命題が成り立つとしたときに，$l = k+1$ でも成り立つことを示す．今，入力テーブルの属性数 P は $k+1$ 以上であるとする．以下の 2 つの場合に分けて証明する．

（場合 1：$\lfloor \frac{S_0}{k+1} \rfloor \geq N_0$ のとき）

I の定義より，$I = 0$ となり，$\lfloor \frac{S_I}{k+1-I} \rfloor = \lfloor \frac{S_0}{k+1} \rfloor$ 個が最大ブロックサイズであり，そのブロック数を持つ解が存在することを示す．

S_0 個のタプルからなるテーブルにおいて，単純 $(k+1)$-多様性を保持させる場合，ブロック数が $\lfloor \frac{S_0}{k+1} \rfloor$ より大きくなると，ブロックに振り分けられたタプル数が k もしくはそれ以下のブロックが存在することになる．したがって，単純 $(k+1)$-多様性を保持した解でブロック数が $\lfloor \frac{S_0}{k+1} \rfloor$ より大きなものは存在しないといえる．よって，ブロック数の上界は $\lfloor \frac{S_0}{k+1} \rfloor$ である．

場合 1 のとき，前述の上界のブロック数のブロックを構成できることを示す．実際に，注目属性値の出現頻度の多いタプルグループから順に 1 つずつタプルを $\lfloor \frac{S_0}{k+1} \rfloor$ 個のブロックに振り分けて考えると，すべての注目属性値の出現頻度が $\lfloor \frac{S_0}{k+1} \rfloor$ 以下であることから，同一のブロック中に同じ注目属性値を持つタプルが振り分けられることはない（**図 2** (b) 参照）．したがって，ブロック数 $\lfloor \frac{S_0}{k+1} \rfloor$ で単純 $(k+1)$-多様性を保持する解が存在するといえる．よって，命題が成り立つ．

（場合 2：$\lfloor \frac{S_0}{k+1} \rfloor < N_0$ のとき）

帰納的前提として，入力データテーブル PT' の注目属性値の種類数 P' が $P' \geq k$ のとき，テーブル PT' を k-多様化させる場合，補題 2 の命題が真であるとする．

入力テーブル PT において出現頻度が一番多い N_0 であるタプルグループを取り除き，注目属性値の種類数 P' が $P' = P-1$ であるテーブルを PT'（$N'_i = N_{i+1}, S'_i = S_{i+1}, 1 \leq i \leq P'$）とする．また，$P \geq k+1$ より $P' \geq k$ であるので前提条件から，$I' = \min\{i | \lfloor \frac{S'_i}{k-i} \rfloor \geq N'_i\}$ としたとき，テーブル PT' を k-多様化した場合のブロック数の上界を，$\lfloor \frac{S'_{I'}}{k-I'} \rfloor$ とおくことができ，かつ，その上界のブロック数を持つ単純 k-多様化問題の許容解が存在する．このとき，テーブル PT の $(k+1)$-多様化の任意の解は，テーブル PT' の k-多様化問題の許容解のブロック数上界と同じブロック数上界を持つことを示す．テーブル PT の $(k+1)$-多様化の解は，テーブル PT' の解のブロックに出現頻度が N_0 のタプルグループを加えたと見ることができる．もし，テーブル T' のブロック数を $\lfloor \frac{S'_{I'}}{k-I'} \rfloor$ より大きくすると，テーブル PT' の変換後のテーブルの k-多様性が崩れてしまう．したがって，出現頻度が N_0 であるタプルグループを加えて $(k+1)$-多様性のテーブルにしようとした場合もブロック数上界は $\lfloor \frac{S'_{I'}}{k-I'} \rfloor$ になる．場合 2 の条件より，$I = I' + 1$ となり $\lfloor \frac{S'_{I'}}{k-I'} \rfloor = \lfloor \frac{S_{I'+1}}{(k+1)-(I'+1)} \rfloor = \lfloor \frac{S_I}{(k+1)-I} \rfloor$．また，$N_0 \geq N'_0$ より，$N_0 > \lfloor \frac{S_{I'}}{k-I'} \rfloor$ がいえるので，テーブル T' の k-多様化の解のすべてのブロックに 1 個以上の出現頻度が N_0 であるタプルを加えることができ，テーブル T の最大ブロック数 $\lfloor \frac{S_I}{(k+1)-I} \rfloor$ を持つ単純 $(k+1)$-多様化問題の許容解を構成できる（図 2 (c) 参照）．よって，この場合も補題 2 は真であるといえる．

以上の帰納的証明とエントロピー l-多様化問題，単純 l-多様化問題とも各ブロックに少なくとも l 個の異なる注目属性値を持つタプルが必要なことから，ブロック数が $\lfloor \frac{S_I}{l-I} \rfloor$ より大きな l-多様性問題の許容解は存在しないといえる．また，単純 l-多様化問題の場合，上記議論に従い帰納的にブロック数 $\lfloor \frac{S_I}{l-I} \rfloor$ を持つ解を構成できる．よって，任意の l（$\leq P$）について補題 2 が成立する．　□

3.3　エントロピー l-多様化問題での最大ブロックサイズの下界

補題 2 をエントロピー l-多様化基準での性質を述べるために拡張することを考える．エントロピー l-多様化基準では，各ブロックは，少なくとも l 種類の異なる注目属性値を含んでいなければならないが，すべてのブロックの注目属性値の種類を均一にすることが必ずしもブロック数を最大にすることにはつながらない．たとえば，注目属性値が l 種類のみのサイズの小さなエントロピー条件を満たすブロックを多くつくり，l' 種類（$l' > l$）の大きなブロックを少量作成することでエントロピー l-多様化が達成できる可能性がある．そこで，本稿では，エントロピー l-多様化基準を満たす最も大きなブロックのサイズ，つまり，最大ブロックサイズの下界 α のみについて補題 2 を拡張する定理を示す．以下にまず，定理で用いる語を定義する．

定義 5　エントロピー多様さ関数
データテーブル PT のあるブロック分割 $B = \{b_1, b_2, \ldots\}$ が与えられたときに B の多様さを返す関数 $Entropy : \mathcal{B} \to R^+$ を以下に定義する．ここで，$\mathcal{B}$ は任意のブロック分割からなる集合，R^+ は 0 以上の実数の集合とする．また，関数 $P(b, s)$ はブロック b に属し，かつ，タプルで注目属性値の値が s であるタプルの数をブロック b のタプル数で割った出現頻度を返すとする．ここでは関数 $f(x) = x \log x$ は $x = 0$ のとき $f(0) = 0$ と定義する．

$$Entropy(B) = \min\{-\sum_{s \in S} P(b, s) \log(P(b, s)) | b \in B\}$$

定理 1　注目属性値を P 種類持つ入力データテーブル PT において，$\alpha \leq N$ なる自然数 α，最大ブロックのサイズが α であるデータテーブル PT の任意のブロック分割 B では，ブロック分割 B のエントロピー多様さ関数値 $Entropy(B)$ は以下の不等式を満たす．

$$Entropy(B) \leq -\sum_{0 \leq i < I} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \frac{\log \alpha}{S_0} S_I.$$

ここで，$I = \min\{i | N_i \leq S_0/\alpha, 0 \leq i < P\}$ とする．

[証明]　最大ブロックサイズ α に関する数学的帰納法により証明する．

(基本ステップ)
$\alpha = 1$ のとき，ブロックに含まれるすべてのタプル数が 1 である任意のブロック分割 B のエントロピー多様さ関数値 $Entropy(B)$ は定義より 0 である．一方，命題での I はすべての注目属性値の種類 i について，$N_i \leq S_0$ であることより，$I = 0$ を満たす．よって，命題の不等式の右辺は 0 となり，命題が成り立つ．

(帰納ステップ)
任意の自然数 q（$q > 1$）において $\alpha \leq q$ を満たすブロックサイズ α で命題が成り立つと仮定する．ブロック分割 B の最大ブロックサイズが $q+1$ であるとき，B の中でブロックサイズが $q+1$ のブロックのみを集めた集合を $B(q+1)$ とする．ブロック集合 $B - B(q+1)$ で表されるサイズが q 以下のブロック集合を $B_{\leq}(q)$ で表すとする．また，ブロック集合 $B(q+1)$ に属するブロックに含まれるタプルのみによって構成されるデータテーブルを T_A（$T_A = \cup_{b \in B(q+1)} b$）とする．同様にブロック集合 $B_{\leq}(q)$ に対するデータテーブル T_B を定義する．このとき，データテーブル T_A が $T_A = PT$ を満たす場合とそうでない場合について考察する．以降では，$I = \min\{i | N_i \leq S_0/(q+1), 0 \leq i < P\}$ とする．

(場合 1：$T_A = PT$ のとき)
ブロック分割 B に属するすべてのブロックはサイズ $(q+1)$ を持つ．すべてのブロックが同じサイズであるとき，最小エントロピーを最大化する最も理想的で無駄のないブロック分割はすべてのブロックが同じ分布になるように均等に各種類のタプルを分布させることである．$0 \leq i < I$ の種類 i ではすべてのブロックに 1 個より多いタプルを割り当てることができる．一方，$I \leq i < P$ を満たす種類では，実際には許容解では各タプルを整数単位でしか分布させることができないため，ブロックに属することのできる種類の数がブロックサイズによって変化する．つまり $I \leq i < P$ を満たす種類 i の各タプルがブロックのエントロピーに寄与できる最大値はブロックサイズ $(q+1)$ のブロックで同じ種類のタプルがブロックにないタプルによってもたらされるエントロピーの値 $-\frac{1}{q+1} \log \frac{1}{q+1}$ である．よって，$I \leq i < P$ を満たすすべての種類のタプル集合によって各ブロックにもたらされるエントロピーの平均値の上界は $-\sum_{I \leq i < P} \frac{N_i(q+1)}{S_0} (\frac{1}{q+1} \log \frac{1}{q+1}) = \frac{S_I}{S_0} \log(q+1)$ となる．よって，以下の不等式が成り立つ．

$$Entropy(B) \leq -\sum_{0 \leq i < I} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \frac{\log(q+1)}{S_0} S_I.$$

(場合 2：$T_A \neq PT$ のとき)
データテーブル T_A と T_B が $|T_A| = \gamma |T_B|$ を満たすとする．ここで，$\gamma > 0$. データテーブル PT をタプル数が $\gamma : 1$ となるよう各種類 i（$0 \leq i < P$）のタプルを均等に分配（1 つのタプルを実数の任意の量の部分に分割できると仮定する）した 2 つのデータテーブルを T_{A0}，T_{B0} とし，それぞれのテーブルの各種類 i のタプル数を N_{Ai}，N_{Bi}（N_{Ai}，$N_{Bi} \in R^+$）で表す．データテーブル T_A の各種類 i のタプル数を均等データテーブル T_{A0} のタプル数を用い $N_{Ai} - d_i$ で表せるよう d_i（$d_i \in R$，ここで R は実数の集合とする）を定めると，データテーブル T_B の同種類 i のタプル数は $N_{Bi} + d_i$ と表される．また，すべての種類 i の値 d_i の総和は $\sum_{0 \leq i < P} d_i = 0$ を満たす．このとき，データテーブル T_A のブロック集合 $B(q+1)$ とデータテーブル T_B のブロック分割 $B_{\leq}(q)$ のエントロピーの関係を考察する．ここで，定義よりブロック集合 $B_{\leq}(q)$ はブロックサイズが q 以下のブロックしか含まれない．また，関数 $Entropy$ の定義より，$Entropy(B) \leq \min\{Entropy(B(q+1)), Entropy(B_{\leq}(q))\}$ が成り立つ．

帰納法の仮定より以下の不等式が成り立つ．ここで $I' = \{i | N_{Bi} + d_i \leq |T_B|/q, 0 \leq i < P\}$ とする．

$$Entropy(B_{\leq}(q)) \leq -\sum_{i \notin I'} \frac{N_{Bi}+d_i}{|T_B|} \log \frac{N_{Bi}+d_i}{|T_B|} + \frac{\log q}{|T_B|} \sum_{i \in I'} (N_{Bi}+d_i)$$

このとき，補題 A がいえる（証明は付録に記載）．

[補題 A]　ある実数 η において，データテーブル T_B が以下を満たすとする．

$$\eta \leq -\sum_{i \notin I'} \frac{N_{Bi}+d_i}{|T_B|} \log \frac{N_{Bi}+d_i}{|T_B|} + \frac{\log q}{|T_B|} \sum_{i \in I'} (N_{Bi}+d_i)$$

このとき，任意の集合 $\mathcal{I}$（$\mathcal{I} \subseteq \{0,1,\ldots,P-1\}$）において以下の不等式が成り立つ．

$$\eta \leq -\sum_{i \notin \mathcal{I}} \frac{N_{Bi}+d_i}{|T_B|} \log \frac{N_{Bi}+d_i}{|T_B|} + \frac{\log q}{|T_B|} \sum_{i \in \mathcal{I}} (N_{Bi}+d_i) \qquad \square$$

よって，補題 A より以下の不等式が成り立つ．

$$Entropy(B_{\leq}(q)) \leq -\sum_{0 \leq i < I} \frac{N_{Bi}+d_i}{|T_B|} \log \frac{N_{Bi}+d_i}{|T_B|} + \frac{\log q}{|T_B|} \sum_{I \leq i < P} (N_{Bi}+d_i)$$

$$Entropy(B_{\leq}(q)) + \sum_{0 \leq i < I} \frac{N_{Bi}}{|T_B|} \log \frac{N_{Bi}}{|T_B|}$$
$$\leq -\sum_{0 \leq i < I} \left(\frac{N_{Bi}+d_i}{|T_B|} \log \frac{N_{Bi}+d_i}{|T_B|} - \frac{N_{Bi}}{|T_B|} \log \frac{N_{Bi}}{|T_B|}\right) + \frac{\log q}{|T_B|} \sum_{I \leq i < P} (N_{Bi}+d_i) \quad (2)$$

今，$Entropy(B) = -\sum_{0 \leq i < I} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \Delta E = -\sum_{0 \leq i < I} \frac{N_{Bi}}{|T_B|} \log \frac{N_{Bi}}{|T_B|} + \Delta E$ と表すことができるよう，ΔE を定義する．このとき，式 (2) は以下のように変形できる．

$$-\sum_{0 \leq i < I} \left(\frac{N_{Bi}+d_i}{|T_B|} \log \frac{N_{Bi}+d_i}{|T_B|} - \frac{N_{Bi}}{|T_B|} \log \frac{N_{Bi}}{|T_B|}\right) \geq \Delta E - \frac{\log q}{|T_B|} \sum_{I \leq i < P} (N_{Bi}+d_i) \quad (3)$$

一方，以下の補題 B が成り立つ（証明は付録に記載）．

[補題 B]　ある実数 β において，データテーブル T_B が以下を満たすとする．

$$-\sum_{0 \leq i < I} \left(\frac{N_{Bi}+d_i}{|T_B|} \log \frac{N_{Bi}+d_i}{|T_B|} - \frac{N_{Bi}}{|T_B|} \log \frac{N_{Bi}}{|T_B|}\right) \geq \beta$$

このとき，データテーブル T_A は以下の不等式を満たす．

$$-\sum_{0 \leq i < I} \left(\frac{N_{Ai}-d_i}{|T_A|} \log \frac{N_{Ai}-d_i}{|T_A|} - \frac{N_{Ai}}{|T_A|} \log \frac{N_{Ai}}{|T_A|}\right) \leq -\frac{\beta}{\gamma} \qquad \square$$

上記の補題 B と式 (3) より以下の不等式が成り立つ．

$$-\sum_{0 \leq i < I} \left(\frac{N_{Ai}-d_i}{|T_A|} \log \frac{N_{Ai}-d_i}{|T_A|} - \frac{N_{Ai}}{|T_A|} \log \frac{N_{Ai}}{|T_A|}\right) \leq -\frac{1}{\gamma}\left(\Delta E - \frac{\log q}{|T_B|} \sum_{I \leq i < P} (N_{Bi}+d_i)\right) \quad (4)$$

一方，場合 1 と同様の議論により以下の不等式が成り立つ．ここで $I'' = \{i | N_{Ai} \leq |T_A|/(q+1), 0 \leq i < P\}$ とする．

$$Entropy(B(q+1)) \leq -\sum_{i \notin I''} \frac{N_{Ai}-d_i}{|T_A|} \log \frac{N_{Ai}-d_i}{|T_A|} + \frac{\log(q+1)}{|T_A|} \sum_{i \in I''} (N_{Ai}-d_i)$$

上記の不等式，および，補題 A より以下がいえる．

$$Entropy(B(q+1)) \leq -\sum_{0 \leq i < I} \frac{N_{Ai}-d_i}{|T_A|} \log \frac{N_{Ai}-d_i}{|T_A|} + \frac{\log(q+1)}{|T_A|} \sum_{I \leq i < P} (N_{Ai}-d_i) \quad (5)$$

また，ΔE の定義 $\Delta E \leq Entropy(B(q+1)) + \sum_{0 \leq i < I} \frac{N_{Ai}}{|T_A|} \log \frac{N_{Ai}}{|T_A|}$，式 (5) より以下の不等式が成り立つ．

$$\Delta E \leq -\sum_{0 \leq i < I} \left(\frac{N_{Ai}-d_i}{|T_A|} \log \frac{N_{Ai}-d_i}{|T_A|} - \frac{N_{Ai}}{|T_A|} \log \frac{N_{Ai}}{|T_A|}\right) + \frac{\log(q+1)}{|T_A|} \sum_{I \leq i < P} (N_{Ai}-d_i) \quad (6)$$

式 (4) を式 (6) に代入すると以下の不等式が成り立つ．

$$\Delta E \leq -\frac{1}{\gamma}\left(\Delta E - \frac{\log q}{|T_B|} \sum_{I \leq i < P} (N_{Bi}+d_i)\right) + \frac{\log(q+1)}{|T_A|} \sum_{I \leq i < P} (N_{Ai}-d_i)$$
$$\frac{1+\gamma}{\gamma} \Delta E \leq \frac{1}{\gamma} \frac{\log q}{|T_B|} \sum_{I \leq i < P} (N_{Bi}+d_i) + \frac{\log(q+1)}{|T_A|} \sum_{I \leq i < P} (N_{Ai}-d_i)$$
$$\Delta E \leq \frac{1}{\gamma+1} \frac{\log q}{|T_B|} \sum_{I \leq i < P} (N_{Bi}+d_i) + \frac{\gamma}{\gamma+1} \frac{\log(q+1)}{|T_A|} \sum_{I \leq i < P} (N_{Ai}-d_i)$$
$$= \frac{\log q}{S_0} \sum_{I \leq i < P} (N_{Bi}+d_i) + \frac{\log(q+1)}{S_0} \sum_{I \leq i < P} (N_{Ai}-d_i)$$

$$\leq \frac{\log(q+1)}{S_0} \sum_{I \leq i < P} (N_{Bi} + d_i + N_{Ai} - d_i) = \frac{\log(q+1)}{S_0} \sum_{I \leq i < P} N_i$$

よって，以下の不等式が成り立つ．

$$Entropy(B) = -\sum_{0 \leq i < I} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \Delta E \leq -\sum_{0 \leq i < I} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \frac{\log(q+1)}{S_0} S_I$$

□

系 1　入力データテーブル PT，$l \leq P$ なる自然数 l において，

$$I = \min\{i | (-\sum_{0 \leq j < i} \frac{N_j}{S_0} \log \frac{N_j}{S_0} + \frac{S_i}{S_0} \log \lfloor \frac{S_0}{N_i} \rfloor \geq \log l \wedge 0 \leq i < P) \vee (i = P)\}$$

と定義した場合，$I < P$ なら入力データテーブル PT に対する最大ブロックサイズが下記の α より小さなエントロピー l-多様化問題の許容解は存在しない．ここで x は対数関数の底の値とする．

$$\alpha = \lceil x^{\frac{S_0}{S_I}(\log l + \sum_{0 \leq j < I} \frac{N_j}{S_0} \log \frac{N_j}{S_0})} \rceil$$

$I = P$ ならテーブル PT に対するエントロピー l-多様化問題の許容解は存在しない．　□

［適用例］　注目属性値の種類数 $P = 5$，各種類のタプル数 N_0，N_1，N_2，N_3，N_4 をそれぞれ 10，8，7，3，2 とするデータテーブル PT_1 を考える．このとき $l = 3$ の多様性を満たすブロック分割について考察する．単純 3-多様性を満たす任意のブロック分割のブロック数の最大値は補題 1 より 10 となり，最大ブロックサイズの下界は $30/10 = 3$ となる．一方，エントロピー 3-多様性を満たすブロック分割の最大ブロックサイズに関しては，系 1 より $I = 0$，$\alpha = \lceil x^{\frac{S_0}{S_I}(\log l + \sum_{0 \leq j < I} \frac{N_j}{S_0} \log \frac{N_j}{S_0})} \rceil = 2^{\log 3}$ となり 3 の下界を持つ．ここでは log 関数の底を 2 とした．データテーブル PT_1 の場合，補題 1 の証明での議論したブロック分割方法によって最大ブロックサイズ 3 のエントロピー 3-多様性を満たすブロック分割を構成できるため，上記の下界は上界と一致する．

注目属性値の種類数 $P = 5$，各種類のタプル数を N_0，N_1，N_2，N_3，N_4 をそれぞれ 50，25，15，7，3 とするデータテーブル PT_2 を考える．このとき $l = 3$ の多様性を満たすブロック分割について考察する．単純 3-多様性を満たすブロック分割のブロック数の最大値は補題 2 より 25 となり，最大ブロックサイズの下界は $100/25 = 4$ となる．一方，系 1 より $I = 2$，$S_I = 25$ となりエントロピー 3-多様性を満たすには $\alpha = \lceil x^{\frac{S_0}{S_I}(\log l + \sum_{0 \leq j < I} \frac{N_j}{S_0} \log \frac{N_j}{S_0})} \rceil = \lceil 2^{4(\log 3 + \frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4})} \rceil = \lceil 5.06 \rceil = 6$ となり最大ブロックサイズは 6 以上となる．テーブル PT_2 の場合，補題 2 の証明で議論したブロック分割方法によるブロック分割はそのエントロピー多様さの値が約 1.371 となり，エントロピー 3-多様性は達成できていない．　□

4. 提案指標の評価

4.1 評価実験

データを入力として与え，k-匿名化問題と l-多様化問題の最適歪曲度の下界をシミュレーション実験で算出した．入力に用いたデータは，人工データであるランダムデータ（本稿の例にあげた医療データに数種類の整数値属性を加えたもの）2 種類（*data*1，*data*2），ベンチマークデータとして *University of California, Irvine* の *KDD*（*Knowledge Discovery in Databases*）アーカイブ（http://kdd.ics.uci.edu）からの *coil*2000 (ticdata2000.txt) の保険会社のデータ（*data*3，*data*4）と *Japanese Vowels* (ae.test) データ（data5）と *IPMUS* 国勢調査（ipmus9.la.97）データからタプル数 10000，属性数 11 を抜き出したデータ（*data*6）を使用した．また，各入力テーブルにおける注目属性として，*data*3 は属性 *MOSTYPE*（顧客サブタイプ），*data*4 は属性 *MINKGEM*（平均所得），*data*5 は第 12 番目の属性（LPC 系列），*data*6 は属性 *nmothers* と設定した．各データの属性の値一般化階層は図 1 の値一般化階層の例と同様に属性値を文字列と見なし，最後尾の文字より 1 文字ずつ先頭方向に* に置き換える単純な階層とした．

各データを k-匿名化，l-多様化するのに必要である最適歪曲度の下界を**表 3** に示す．表中での NA は系 1 の適用によって該当データの l-多様化問題の許容解が存在しないことが判別されたことを示す．また，最大 BN は単純 l-多様化問題の最大ブロック数，最大 BS_s は単純 l-多様化問題としての最大ブロック数の下界，最大 BS_e はエントロピー多様化問題としての最大ブロック数の下界を示す．また，得られた最適歪曲度の下界と l-多様性アルゴリズム *DiverDIS*（エントロピー多様化）を適用して得られた歪曲度の上界，k-匿名性アルゴリズム *MinDIS* を適用して得られた歪曲度の上界との近似比も示す．

4.2 考　　察

表 3 より k-匿名化問題に対する発見的手法 $MinDIS$ を適用して得られた歪曲度の上界は下界との比により，data1 を除いては 1.7 倍以下であり，最適解との比も 1.7 倍以下とい

表 3　l-多様化と k-匿名化の下界に関する比較
Table 3　Comparison of lower bounds between l-diversification and k-anonymization.

Data			l-多様性							k-匿名性		
名前	タプル数	非注目属性数	l	歪曲度の下界	歪曲度の上界	上界/下界	最大 BN	最大 BS_s	最大 BS_e	k	歪曲度の下界	上界/下界
data1	1000	6	2	0.0089	0.039	4.38	500	2	2	2	0.0086	3.37
			4	0.0269	0.125	4.64	250	4	4	4	0.0128	3.67
			8	0.1493			108	10	11	8	0.0555	2.49
data2	10000	6	4	0.2670	0.413	1.55	2500	4	4	4	0.2666	1.70
			8	0.2883	0.468	1.62	1250	8	8	8	0.2876	1.45
			10	0.2943	0.484	1.64	1000	10	10	10	0.2930	1.62
data3	5822	85	2	0.1667	0.563	3.38	2911	2	2	2	0.0497	1.59
			4	0.2350	0.819	3.49	1456	4	4	4	0.0815	1.66
			8	0.2780			716	9	9	8	0.1186	1.15
data4	5822	85	2	0.2337	0.631	2.70	2911	2	2	2	0.0504	1.61
			4	0.2877	0.829	2.88	1297	5	7	4	0.0828	1.64
			8	0.3505	NA	NA	540	11	NA	8	0.1189	1.15
data5	5687	11	2	0.5743	0.589	1.03	2844	2	2	2	0.5743	1.04
			4	0.5947	0.718	1.21	1422	4	4	4	0.5947	1.28
			8	0.6078	0.773	1.27	711	8	8	8	0.6078	1.29
data6	10000	10	2	0.0950	0.405	4.26	5000	2	2	2	0.0025	1.20
			4	0.1182	0.460	3.89	1716	6	8	4	0.0172	1.40
			8	0.1426	NA	NA	627	16	NA	8	0.0358	1.23

える．これより，k-匿名化問題に対しては最適歪曲度に近い上界，下界を求めていることが分かる．一方，l-多様化問題に対しては発見的手法 $DiverDIS$ を適用して得られた歪曲度の上界は下界との比において data5 を除いて 1.5〜4.64 倍となり上界と下界に差がある．上記は下界か，上界のどちらかもしくは両方が最適歪曲度と離れている可能性を示している．また，最大ブロック数の下界を調べることで許容解が存在しない場合を一部判定できている．

k-匿名化した場合の最適歪曲度の下界を 1 としたときの，同じ入力テーブルでの l-多様化問題の最適歪曲度の下界，$MinDIS$ 導出解，$DiverDIS$ 導出解とのデータ歪曲度の比率をグラフ化したものを**図 3** に示す．図 3 の結果から，l-多様化の下界と k-匿名化の下界においてデータ歪曲度に差があるデータは，発見的手法で得られた解どうしにもデータ歪曲度の差が大きく見られた．また，そのようなデータは k-匿名化アルゴリズム $MinDIS$ が導出した解のデータ歪曲度より，l-多様化問題の最適データ歪曲度の下界のほうが高いという傾向も見て取れた．よって，これらの傾向を示すデータは l-多様化の適性が低いと判断できる．一方で，data5 のように l-多様化問題でも k-匿名化問題でも最適解に近い解を発見的手法

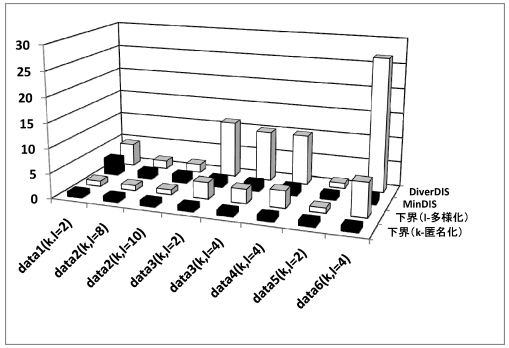


図 3　下界と発見的手法のデータ歪曲度による比較
Fig. 3　Comparison between lower bounds of data distortion and upper bounds outputted by heuristic algorithms.

で求めている例もある．data5 ではほとんどすべての注目属性値が異なっており，k-匿名化問題と l-多様化問題がほぼ同程度の難しさを持っている例である．

導出した下界，上界の妥当性をさらに詳しく調べるため，k-匿名化問題と比較したときの l-多様化問題の難しさを表現することを目的として l-多様化問題の最適歪曲度の下界と k-匿名化問題の最適解の上界（$MinDIS$ による解）を比較する．**表 4** ではそれぞれの入力データに対する l-多様化問題の最適歪曲度の下界を k-匿名化問題の最適化の上界で割った値を記している．ここでの匿名化のレベル k は比較対象として妥当と考えられる表 3 の最大 BS_e を用いた．これらの比は 2 つの問題の難易度の比を最適歪曲度の比として表したときの k-匿名化問題の最適歪曲度に対する l-多様化問題の最適歪曲度の比の下界となる．つまり，data6 では 2-多様化問題は 2-匿名化問題の 16 倍以上も歪曲しなければ解くことができないことを示している．このことは data6 では 2-匿名化問題の解ではほとんど注目属性値に関して多様性がないことを暗示しており，同種攻撃が成立しやすい状況であることが分かる．一方，data4 では先に述べたように l-多様化問題は k-匿名化問題と同程度の難易度を持っているが，表中の比も 1 未満の値となっている．

l-多様化の適性を k-匿名化に必要な歪曲度と l-多様化に必要な歪曲度の比で表現したとすると k-匿名化に比べ問題自体が難しい場合があることが分かった．しかしながら，適性が低いデータほど，k-匿名化データでは同種攻撃や背景知識攻撃に対する耐性が低いため，プライバシ要求を満たすことが重要なデータかをデータ保有者が確認する必要がある．データ

表 4　l-多様化下界と k-匿名化上界のデータ歪曲度比率
Table 4　Comparison between lower bounds of data distortion degrees of l-diversification and upper bounds of data distortion degrees of k-anonymization.

Data	l-多様性		k-匿名性		l-多様性の下界/
名前	l	歪曲度の下界	k（最大 BS_e）	歪曲度の上界	k-匿名性の上界
data1	2	0.0089	2	0.029	0.3068
	4	0.0269	4	0.047	0.5723
	8	0.1493	11	0.123	1.2134
data2	4	0.2670	4	0.452	0.5907
	8	0.2883	8	0.469	0.6147
	10	0.2943	10	0.475	0.6200
data3	2	0.1667	2	0.079	2.1101
	4	0.2350	4	0.135	1.7407
	8	0.2780	9	0.139	2.0000
data4	2	0.2337	2	0.081	2.8852
	4	0.2877	7	0.139	2.0698
	8	0.3505	NA	-	-
data5	2	0.5743	2	0.599	0.9588
	4	0.5947	4	0.764	0.7784
	8	0.6078	8	0.762	0.7976
data6	2	0.0950	2	0.003	16.667
	4	0.1182	8	0.044	2.6863
	8	0.1426	NA	-	-

保有者の確認のうえで，妥当な保護レベルを保持した公開データを作成できない場合には公開を断念せざるをえない．

5. 応　用　例

前章で示した l-多様化適性評価を取り入れたプライバシ保護システムの構成案を**図 4** に示す．提案システムでは入力データについて l-多様化が可能かどうか，可能な場合にはどのレベルの匿名性保持データと歪曲度を比較することが妥当かを最大ブロックサイズの下界を求めることで判断する．そのうえで，k-匿名性を保持させた場合のデータ歪曲度の上界（発見的アルゴリズムによる）と l-多様性を保持させるために必要なデータ歪曲度の下界の比 $ratio$ を求め適性評価を行う．たとえば，$ratio > 2$ であった場合には l-多様化適性が低いと判断しデータ提供者がデータのプライバシ保護レベルを維持する必要があるか，または多様性が低い匿名化データの公開に危険がないかを判断することを支援する情報を提供する．非常に偏った分布のデータである場合，同種攻撃の対象となっても問題のない注目属性値が

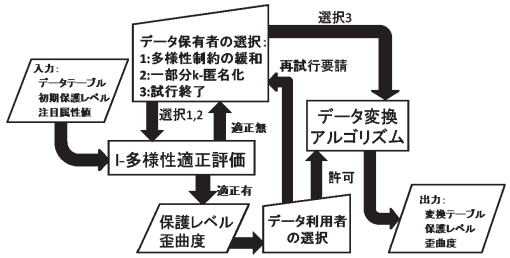


図 4　プライバシ保護システム
Fig. 4　Privacy Protection System.

多く分布しているのかそうでないかの確認を求める．データ保有者がデータを確認後，プライバシ保護要求が許す限りにおいて選択できる妥協案を提示する．プライバシ保護要求が妥協を許さない場合には適性は低いがそのままのプライバシレベルで l-多様化を行うようシステムから提案する．また，データ利用者に対しては l-多様化後のデータ歪曲度が高くなる要因がデータによるものなのか，データ変換方法によるものなのかの判断を支援できるよう歪曲度の下界を示す．データ変換アルゴリズムによりさらに低い歪曲度の変換テーブルを作成できる可能性があると利用者が判断した場合には，システムに再試行を要請する．

6. おわりに

本稿では，入力データの l-多様化の適性の評価に用いることを目的に 3 つの指標を提案した．提案指標として，l-多様化問題の最適歪曲コストの下界を導出した．さらに問題の許容解でのブロック数，ブロックサイズに注目し，ブロック数の上界，最大ブロックサイズの下界の導出式を定義した．シミュレーション実験により，導出指標の l-多様化適性の判断指標としての有効性を確認した．

今後の課題として文献 4) で扱われている解析技術（たとえば，検索や分類）に合わせた一般化階層を決定する手法を組み込んだ場合の提案指標のさらなる深い評価があげられる．将来的には，データ歪曲度が低くなるよう，属性値の持つ意味を考慮し l-多様化と k-匿名化を組み合わせ，かつ，文献 6) で扱われている数値変換であるような注目属性での多様性を考慮できる，データ管理者，使用者にとって高価値といえる出力テーブルを導出するプライバシ保護システムの実現を目指す．

参 考 文 献

1) Aggarwal, G., Feder, T., Kenthapadi, K., Motowani, R., Panigrahy, R., Thomas, D. and Zhu, A.: Approximation algorithm for k-anonymity, *Journal of Privacy Technology*, Paper no.20051120001 (2005).
2) Fung, B., Wang, K., Chen, R. and Yu, P.: Privacy-preserving data publishing: A survey on recent developments, *ACM Computing Surveys*, Vol.42, No.4 (2010).
3) LeFevre, K., DeWitt, D.J. and Rawakrishnan, R.: Incognito: Efficient full-domain k-anonymity, *Proc. ACM SIGMOD International Conference on Management of Data 2005*, pp.49–60 (2005).
4) LeFevre, K., DeWitt, D.J. and Rawakrishnan, R.: Workload-aware anonymization techniques for large-scale datasets, *ACM Trans. Database Systems*, Vol.33, No.3, Article 17 (2008).
5) Li, N., Li, T. and Venkatasubramanian, S.: t-Closeness: Privacy beyond k-anonymity and l-diversity, *Proc. 21st IEEE International Conference on Data Engineering* (*ICDE*) (2007).
6) Li, J., Tao, Y. and Xiao, X.: Preservation of proximity privacy in publishing numerical sensitive data, *Proc. ACM SIGMOD Conference on Management of Data 2008*, pp.473–486 (2008).
7) Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M.: l-Diversity: Privacy beyond k-anonymity, *Proc. 22nd IEEE International Conference on Data Engineering* (*ICDE*) (2006).
8) Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: l-Diversity: Privacy beyond k-anonymity, *ACM Trans. Knowledge Discovery from Data*, Vol.1, No.1, Article 3 (2007).
9) 村本俊祐，上土井陽子，若林真一：k-匿名性を利用したデータ一般化によるプライバシー保護，DEWS2007 論文集 (2007).
10) 村本俊祐，上土井陽子，若林真一：データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム，*DBSJ Letters*, Vol.6, No.1, pp.97–100 (2007).
11) 村本俊祐，上土井陽子，若林真一：背景知識を用いた推測を困難にしデータ歪曲度を極小化するプライバシー保護手法，DEWS2008 論文集 (2008).
12) Park, H. and Shim, K.: Approximate algorithms for k-anonymity, *Proc. ACM SIGMOD Internatinal Conference on Management of Data 2007*, pp.67–78 (2007).
13) Samarati, P.: Protecting respondents' identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering*, Vol.13, No.6, pp.1010–1027 (2001).
14) Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol.10, No.5, pp.571–588 (2002).
15) Xiao, X., Yi, K. and Tao, Y.: The hardness and approximation algorithms for l-diversity, *Proc. 13th International Conference on Extending Database Technology* (*EDBT2010*), pp.135–146 (2010).

付　　録

[補題 A]　データテーブル PT を P 種類の注目属性値を持つテーブルとし，各種類のタプル数の非増加順に種類を番号付け $0, 1, \ldots, P-1$ とし，各種類のタプル数を N_i（$N_0 \geq N_1 \geq \ldots \geq N_{P-1}$）とする．また，

$$S_i = \sum_{i \leq j < P} (Nj)$$

と S_i（S_0 はタプル総数に対応）を定義する．このとき，ある正整数 α（$1 \leq \alpha \leq S_0$）が与えられたときに種類のインデックスの集合 X は $X = \{i | N_i \leq S_0/\alpha, 1 \leq i < P\}$ を満たすとする．

今，データテーブル PT がある実数 η において以下の不等式を満たすとする．

$$-\sum_{i \notin X} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \frac{\log \alpha}{S_0} \sum_{i \in X} N_i \geq \eta$$

このとき，任意の種類のインデックス集合 $Y \subseteq \{0, \ldots, P-1\}$ において以下の不等式が満たされる．

$$-\sum_{i \notin Y} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \frac{\log \alpha}{S_0} \sum_{i \in Y} N_i \geq \eta$$

[証明]　関数 $F(X)$ と $F(Y)$ を以下のように定義する．

$$F(X) = -\sum_{i \notin X} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \frac{\log \alpha}{S_0} \sum_{i \in X} N_i$$

$$F(Y) = -\sum_{i \notin Y} \frac{N_i}{S_0} \log \frac{N_i}{S_0} + \frac{\log \alpha}{S_0} \sum_{i \in Y} N_i$$

このとき，$i \in X \cap Y$ か $i \notin X \cup Y$ を満たす種類 i については $F(X)$，$F(Y)$ とも種類 i が寄与する値は同じである．

種類 i が $i \in X$ かつ $i \notin Y$ を満たすとき，集合 X の定義より $N_i \leq S_0/\alpha$ が成り立つ．よって，$N_i/S_0 \leq 1/\alpha$ から以下の不等式が成り立つ．

$$\frac{\log\alpha}{S_0}N_i \leq -\frac{N_i}{S_0}\log\frac{N_i}{S_0}$$

したがって，$i \in X$ かつ $i \notin Y$ を満たす種類 i では $F(X)$ へ寄与する値よりも $F(Y)$ へ寄与する値の方が大きい．

種類 i が $i \notin X$ かつ $i \in Y$ を満たすとき，集合 X の定義より $N_i > S_0/\alpha$ が成り立つ．よって，$N_i/S_0 > 1/\alpha$ から以下の不等式が成り立つ．

$$\frac{\log\alpha}{S_0}N_i > -\frac{N_i}{S_0}\log\frac{N_i}{S_0}$$

したがって，$i \notin X$ かつ $i \in Y$ を満たす種類 i では $F(X)$ へ寄与する値よりも $F(Y)$ へ寄与する値の方が大きい．

上記の議論より $F(X) \leq F(Y)$ がいえるので命題が成り立つ．　□

[補題 B]　データテーブル PT は P 種類の注目属性値を持つテーブルとし，各種類のタプル数の非増加順に種類を番号付け $0, 1, \ldots, P-1$ とし，各種類のタプル数を N_i（$N_0 \geq N_1 \geq \ldots \geq N_{P-1}$）とする．

$$S_i = \sum_{i \leq j < P}(Nj)$$

と S_i（S_0 はタプル総数に対応）を定義する．このとき，ある正整数 α（$1 \leq \alpha \leq S_0$）が与えられたときに $I = \min\{i | N_i \leq S_0/\alpha, 1 \leq i < P\}$ を満たすとする．データテーブル PT のある 2 分割 T_A，T_B を考え，$|T_A|/|T_B| = \gamma$ とする．また，データテーブル PT の各注目属性値 i（$0 \leq i < P$）において $N_{Ai} = \gamma N_i/(1+\gamma)$，$N_{Bi} = N_i/(1+\gamma)$ とし，データテーブル T_A の各注目属性値 i（$0 \leq i < P$）のタプル数が $N_{Ai} - d_i$ となるよう d_i を定めると，データテーブル T_B の注目属性値 i のタプル数は $N_{Bi} + d_i$ と表せる．また，このとき，$\sum_{0 \leq i < P} d_i = 0$ が成り立つ．今，データテーブル T_B がある実数 β において以下の不等式を満たすとする．

$$\sum_{0 \leq i < I}\left(-\frac{N_{Bi}+d_i}{|T_B|}\log\frac{N_{Bi}+d_i}{|T_B|} + \frac{N_{Bi}}{|T_B|}\log\frac{N_{Bi}}{|T_B|}\right) \geq \beta$$

このとき，データテーブル T_A において以下の不等式が満たされる．

$$\sum_{0 \leq i < I}\left(-\frac{N_{Ai}-d_i}{|T_A|}\log\frac{N_{Ai}-d_i}{|T_A|} + \frac{N_{Ai}}{|T_A|}\log\frac{N_{Ai}}{|T_A|}\right) \leq -\gamma\beta$$

[証明]　命題の仮定より以下の不等式が成り立つ．

$$\sum_{0 \leq i < I}\left(-\frac{\alpha(N_{Bi}+d_i)}{\alpha|T_B|}\log\frac{\alpha(N_{Bi}+d_i)}{\alpha|T_B|} + \frac{\alpha N_{Bi}}{\alpha|T_B|}\log\frac{\alpha N_{Bi}}{\alpha|T_B|}\right) \geq \beta$$

$$\frac{\log\alpha}{|T_B|}\sum_{0 \leq i < I}(N_{Bi}+d_i-N_{Bi})$$

$$+\frac{1}{\alpha}\sum_{0 \leq i < I}\left(-\frac{\alpha(N_{Bi}+d_i)}{|T_B|}\log\frac{\alpha(N_{Bi}+d_i)}{|T_B|} + \frac{\alpha N_{Bi}}{|T_B|}\log\frac{\alpha N_{Bi}}{|T_B|}\right) \geq \beta$$

$$\frac{1}{\alpha}\sum_{0 \leq i < I}\left(-\frac{\alpha(N_{Bi}+d_i)}{|T_B|}\log\frac{\alpha(N_{Bi}+d_i)}{|T_B|} + \frac{\alpha N_{Bi}}{|T_B|}\log\frac{\alpha N_{Bi}}{|T_B|}\right) \geq \beta - \frac{log\alpha}{|T_B|}\sum_{0 \leq i < I} d_i \quad (7)$$

ここで，$\mathcal{I} = \{i | 0 \leq i < I\}$ とすると，$\mathcal{I}$ の 2 分割 $(\mathcal{I}_+, \mathcal{I}_-)$ を $\mathcal{I}_+ = \{i | i \in \mathcal{I}, d_i > 0\}$，$\mathcal{I}_- = \{i | i \in \mathcal{I}, d_i \leq 0\}$ と定義し，注目属性値の集合 $\mathcal{I}_+$，$\mathcal{I}_-$ に関して以下の値を定義する．

$$A = \sum_{i \in \mathcal{I}_-}\left(\frac{\alpha N_{Bi}}{|T_B|}\log\frac{\alpha N_{Bi}}{|T_B|} - \frac{\alpha(N_{Bi}+d_i)}{|T_B|}\log\frac{\alpha(N_{Bi}+d_i)}{|T_B|}\right)$$

$$B = \sum_{i \in \mathcal{I}_+}\left(\frac{\alpha N_{Bi}}{|T_B|}\log\frac{\alpha N_{Bi}}{|T_B|} - \frac{\alpha(N_{Bi}+d_i)}{|T_B|}\log\frac{\alpha(N_{Bi}+d_i)}{|T_B|}\right)$$

とすると，式 (7) より以下がいえる．

$$\frac{1}{\alpha}(A+B) \geq \beta - \frac{\log\alpha}{|T_B|}\sum_{i \in \mathcal{I}} d_i \quad (8)$$

このとき，以下の式の上限について考える．

$$\sum_{0 \leq i < I}\left(-\frac{N_{Ai}-d_i}{|T_A|}\log\frac{N_{Ai}-d_i}{|T_A|} + \frac{N_{Ai}}{|T_A|}\log\frac{N_{Ai}}{|T_A|}\right)$$

上式は以下のように変形することができる．

$$\sum_{i \in \mathcal{I}}\left(-\frac{\alpha(N_{Ai}-d_i)}{\alpha|T_A|}\log\frac{\alpha(N_{Ai}-d_i)}{\alpha|T_A|} + \frac{\alpha N_{Ai}}{\alpha|T_A|}\log\frac{\alpha N_{Ai}}{\alpha|T_A|}\right)$$

$$= -\frac{\log\alpha}{|T_A|}\sum_{i \in \mathcal{I}} d_i + \frac{1}{\alpha}\left(-\frac{\alpha(N_{Ai}-d_i)}{|T_A|}\log\frac{\alpha(N_{Ai}-d_i)}{|T_A|} + \frac{\alpha N_{Ai}}{|T_A|}\log\frac{\alpha N_{A_i}}{|T_A|}\right)$$

$$= -\frac{\gamma \log \alpha}{|T_B|} \sum_{i \in \mathcal{I}} + \frac{1}{\alpha} \sum_{i \in \mathcal{I}} (-\frac{\alpha(N_{Bi} - \gamma d_i)}{|T_B|} \log \frac{\alpha(N_{Bi} - \gamma d_i)}{|T_B|} + \frac{\alpha N_{Bi}}{|T_B|} \log \frac{\alpha N_{Bi}}{|T_B|})$$

このとき，先に定義した $\mathcal{I}_+$，$\mathcal{I}_-$ において以下の値 C，D を定義する．

$$C = \sum_{i \in \mathcal{I}_-} (\frac{\alpha N_{Bi}}{|T_B|} \log \frac{\alpha N_{Bi}}{|T_B|} - \frac{\alpha(N_{Bi} - \gamma d_i)}{|T_B|} \log \frac{\alpha(N_{Bi} - \gamma d_i)}{|T_B|})$$

$$D = \sum_{i \in \mathcal{I}_+} (\frac{\alpha N_{Bi}}{|T_B|} \log \frac{\alpha N_{Bi}}{|T_B|} - \frac{\alpha(N_{Bi} - \gamma d_i)}{|T_B|} \log \frac{\alpha(N_{Bi} - \gamma d_i)}{|T_B|})$$

今，A，B と C，D の関係について考える．$x = \frac{\alpha N_{Bi}}{|T_B|}$ とすると命題の I の定義より，$i \in \mathcal{I}$ のとき，$x > 1$ が成り立つ．また，関数 $f(x)$ を $f(x) = x \log x$ とするとその導関数は $f'(x) = 1 + \log x$ となる．このとき，以下の 4 つの場合について考える．

［$i \in \mathcal{I}_-$ なる各 i において，$0 \leq \gamma < 1$ のとき］

$x > 1$ のとき $f'(x) > 0$ より，$f(x)$ は増加関数であること，$d_i < 0$ より以下の不等式が成り立つ．

$$x \log x - (x + \frac{\gamma \alpha d_i}{|T_B|}) \log(x + \frac{\gamma \alpha d_i}{|T_B|}) \leq (x - \frac{\gamma \alpha d_i}{|T_B|}) \log(x - \frac{\gamma \alpha d_i}{|T_B|}) - x \log x \quad (9)$$

また，$f'(x)$ が増加関数であることより以下の不等式が成り立つ．

$$\gamma(x \log x - (x + \frac{\alpha d_i}{|T_B|}) \log(x + \frac{\alpha d_i}{|T_B|}) \leq x \log x - (x + \frac{\gamma \alpha d_i}{|T_B|}) \log(x + \frac{\gamma \alpha d_i}{|T_B|}) \quad (10)$$

式 (9)，(10) より以下の不等式が成り立つ．

$$\gamma(x \log x - (x + \frac{\alpha d_i}{|T_B|}) \log(x + \frac{\alpha d_i}{|T_B|})) \leq (x - \frac{\gamma \alpha d_i}{|T_B|}) \log(x - \frac{\gamma \alpha d_i}{|T_B|}) - x \log x$$
$$\gamma A \leq -C$$

［$i \in \mathcal{I}_-$ なる各 i において，$\gamma \geq 1$ のとき］

$x > 1$ のとき $f'(x) > 0$ より，$f(x)$ は増加関数であること，$d_i < 0$ より以下の不等式が成り立つ．

$$x \log x - (x + \frac{\alpha d_i}{|T_B|}) \log(x + \frac{\alpha d_i}{|T_B|}) \leq (x - \frac{\alpha d_i}{|T_B|}) \log(x - \frac{\alpha d_i}{|T_B|}) - x \log x \quad (11)$$

また，$f'(x)$ が増加関数であることより以下の不等式が成り立つ．

$$\gamma((x - \frac{\alpha d_i}{|T_B|}) \log(x - \frac{\alpha d_i}{|T_B|}) - x \log x) \leq (x - \frac{\gamma \alpha d_i}{|T_B|}) \log(x - \frac{\gamma \alpha d_i}{|T_B|}) - x \log x \quad (12)$$

式 (11)，(12) より以下の不等式が成り立つ．

$$\gamma(x \log x - (x + \frac{\alpha d_i}{|T_B|}) \log(x + \frac{\alpha d_i}{|T_B|})) \leq (x - \frac{\gamma \alpha d_i}{|T_B|}) \log(x - \frac{\gamma \alpha d_i}{|T_B|}) - x \log x$$
$$\gamma A \leq -C$$

以上 2 つの場合の考察結果より，$0 \leq \gamma$ において以下の不等式が成り立つ．

$$C \leq -\gamma A \quad (13)$$

［$i \in \mathcal{I}_+$ において，$0 \leq \gamma < 1$ のとき］

$x > 1$ のとき $f'(x) > 0$ より，$f(x)$ は増加関数であること，$d_i \geq 0$ より以下の不等式が成り立つ．

$$x \log x - (x - \frac{\gamma \alpha d_i}{|T_B|}) \log(x - \frac{\gamma \alpha d_i}{|T_B|}) \leq (x + \frac{\gamma \alpha d_i}{|T_B|}) \log(x + \frac{\gamma \alpha d_i}{|T_B|}) - x \log x \quad (14)$$

また，$f'(x)$ が増加関数であることより以下の不等式が成り立つ．

$$(x + \frac{\gamma \alpha d_i}{|T_B|}) \log(x + \frac{\gamma \alpha d_i}{|T_B|}) - x \log x \leq \gamma((x + \frac{\alpha d_i}{|T_B|}) \log(x + \frac{\alpha d_i}{|T_B|}) - x \log x) \quad (15)$$

式 (14)，(15) より以下の不等式が成り立つ．

$$x \log x - (x - \frac{\gamma \alpha d_i}{|T_B|}) \log(x - \frac{\gamma \alpha d_i}{|T_B|}) \leq \gamma((x + \frac{\alpha d_i}{|T_B|}) \log(x + \frac{\alpha d_i}{|T_B|}) - x \log x)$$
$$D \leq -\gamma B$$

［$i \in \mathcal{I}_+$ なる各 i において，$\gamma \geq 1$ のとき］

$x > 1$ のとき $f'(x) > 0$ より，$f(x)$ は増加関数であること，$d_i < 0$ より以下の不等式が成り立つ．

$$x \log x - (x - \frac{\alpha d_i}{|T_B|}) \log(x - \frac{\alpha d_i}{|T_B|}) \leq (x + \frac{\alpha d_i}{|T_B|}) \log(x + \frac{\alpha d_i}{|T_B|}) - x \log x \quad (16)$$

また，$f'(x)$ が増加関数であることより以下の不等式が成り立つ．

$$x \log x - (x - \frac{\gamma \alpha d_i}{|T_B|}) \log(x - \frac{\gamma \alpha d_i}{|T_B|}) \leq \gamma(x \log x - (x - \frac{\alpha d_i}{|T_B|}) \log(x - \frac{\alpha d_i}{|T_B|})) \quad (17)$$

式 (16)，(17) より以下の不等式が成り立つ．

$$(x-\frac{\gamma\alpha d_i}{|T_B|})\log(x-\frac{\gamma\alpha d_i}{|T_B|})-x\log x \le \gamma((x+\frac{\alpha d_i}{|T_B|})\log(x+\frac{\alpha d_i}{|T_B|})-x\log x)$$

$$D \le -\gamma B$$

以上 2 つの場合の考察結果より，$0 \le \gamma$ において以下の不等式が成り立つ．

$$D \le -\gamma B \tag{18}$$

上記の場合分けによる考察結果より，任意の γ（$\gamma \ge 0$）において，式 (13)，(18) が成り立つこと，また，式 (8) より以下の不等式が成り立つ．

$$\begin{aligned}
&\sum_{0\le i<I}(-\frac{N_{Ai}-d_i}{|T_A|}\log\frac{N_{Ai}-d_i}{|T_A|}+\frac{N_{Ai}}{|T_A|}\log\frac{N_{Ai}}{|T_A|})\\
&=-\frac{\log\alpha}{|T_A|}\sum_{0\le i<I}d_i+\frac{1}{\alpha}(C+D)\\
&\le-\frac{\gamma\log\alpha}{|T_B|}\sum_{0\le i<I}d_i-\frac{\gamma}{\alpha}(A+B)\\
&\le-\frac{\gamma\log\alpha}{|T_B|}\sum_{0\le i<I}d_i+\gamma(\frac{\log\alpha}{|T_B|}\sum_{0\le i<I}d_i-\beta)\\
&=-\gamma\beta
\end{aligned}$$

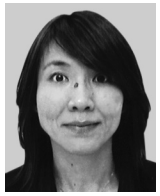
□

(平成 22 年 12 月 20 日受付)
(平成 23 年 4 月 8 日採録)

(担当編集委員　大森 匡)

村本　俊祐

現在，日本電気株式会社．2007 年広島市立大学情報科学部情報工学科卒業．2009 年広島市立大学大学院情報科学研究科修了．在学中はプライバシ保護データ公開に関する研究に従事．



上土井陽子（正会員）

広島市立大学大学院情報科学研究科講師．1994 年広島大学大学院工学研究科博士課程後期修了．博士（工学）．主にデータマイニング，クラスタリングの研究に従事．日本データベース学会，電子情報通信学会，IEEE, ACM，SIAM 各会員．

若林　真一（正会員）

広島市立大学大学院情報科学研究科教授．1984 年広島大学大学院工学研究科博士課程後期修了．工学博士．日本アイ・ビー・エム（株）東京基礎研究所副主任研究員，広島大学工学部助教授を経て，2003 年より現職．主として，VLSI CAD，VLSI 設計，組合せ最適化に関する研究に従事．電子情報通信学会，IEEE，ACM 各会員．