条件付き確率の保守的な推定

菊地　真人[†]　山本　英子[††]　吉田　光男[†]　岡部　正幸[†††]
梅村　恭司[†]

Conservative Estimation for Conditional Probability

Masato KIKUCHI[†], Eiko YAMAMOTO[††], Mitsuo YOSHIDA[†], Masayuki OKABE[†††], and Kyoji UMEMURA[†]

あらまし　本論文では，観測頻度から条件付き確率を推定するという問題に取り組む．条件付き確率の推定は，データマイニングや実際の応用における基本的な操作であり，その推定方法によって手法の正確さが左右されることがある．一般に，確率推定では最ゆう推定値が用いられるが，低頻度に弱いという問題がある．この問題に対処するため，ベイズの枠組みがよく用いられる．ベイズの枠組みでは，データについての事前分布を推定し，事後分布の期待値を用いる．しかし，データをもとに事前分布を推定することは容易ではない．そこで，本論文では，事前分布として何らかの分布を仮定して事後分布の信頼区間を求め，その下限値を用いる手法を提案する．期待値は偏りのない推定値となる一方で，信頼区間の下限値は条件付き確率を保守的に見積もった推定値となる．実験によって，提案手法が低頻度に頑強であることを示す．更に，提案手法は事前分布として一様分布を用いた場合，ベイズの枠組みを用いた手法とほぼ同じ性能を獲得しうることを示す．

キーワード　条件付き確率，信頼区間，下限値，保守的な推定

1. まえがき

観測頻度から条件付き確率を推定することは，データマイニングや実際の応用 [1], [2] における基本的な操作である．二つの事象間に成立する関係を解析する必要があるとき，これらの事象の共起頻度から条件付き確率を推定したい場合がある．例えば，あるトランザクションがアイテム B を含むという条件の下で，そのトランザクションが別のアイテム A を含む可能性を知りたいことがあるかもしれない．このような場合に条件付き確率を推定するが，低頻度に対する扱いがしばしば問題となる．前述した例では，アイテム B を含むトランザクションがごく少数のときに条件付き確率を推定しようとすると，この問題に直面する．

観測頻度やサンプルから確率を推定するとき，最ゆう推定値がよく用いられる．最ゆう推定値は不偏推定量であり，事象を無限に観測したとすると，その事象の発生確率は真の値へと漸近的に収束する．最ゆう推定は低頻度に弱いという欠点がある．ある事象をごくわずかしか観測できなかったとき，事象の発生確率に対する最ゆう推定値は信用できない値となる．

データマイニングでは，観測頻度にしきい値を設け，それ以上の頻度をもつ関係を最ゆう推定値によって求めることがある．これによって，関係を効率良く推定することができる．しかし，ある事象の発生が低頻度であっても，特定の事象とよく共起する場合，低頻度で発生する事象がもう一方の事象の発生を誘発していることがある．頻度にしきい値を設けると，このような関係も推定対象から一律に取り除いてしまうことが問題となる．

確率推定において低頻度の事象を扱うため，ベイズの枠組みがよく用いられる．この枠組みでは，事象の事前分布を仮定し，事象を観測した回数にかかわらず事後分布を計算する．事象を観測した回数が少ないと

[†] 豊橋技術科学大学情報・知能工学系，豊橋市
Department of Computer Science and Engineering, Toyohashi University of Technology, 1–1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, 441–8580 Japan
[††] 岐阜聖徳学園大学経済情報学部，岐阜市
Department of Economics and Information, Gifu Shotoku Gakuen University, 1–38 Nakauzura, Gifu-shi, 500–8288 Japan
[†††] 県立広島大学経営情報学科，広島市
Faculty of Management and Information System, Prefectural University of Hiroshima, 1–1–71 Ujina-Higashi, Minami-ku, Hiroshima-shi, 734–8558 Japan
DOI:10.14923/transinfj.2016DEP0016

き，事後分布の分散は大きくなる．そのため，条件付き確率の事後分布から推定値を決定する場合は注意しなければならない．事前分布を一様分布と仮定して，事後分布の期待値を選択すると，その値はラプラススムージングによる推定値と等しくなる [3]．

本論文では，条件付き確率を推定するために，事後分布の信頼区間を構成し，下限値を用いることを提案する．提案手法は，事前分布として何らかの分布を仮定し，結果を利用するときの適合率に応じて二つの事象間に成立する関係の強さを保守的に推定する．この推定値は事後分布の分散を考慮した値となり，最ゆう推定では扱いにくい低頻度の事象に対しても適切に対処できる．また，提案手法は事前分布として何らかの分布を仮定するが，推定対象となるデータについての事前分布は推定しない．実験において，人工的に生成したデータ集合及び新聞記事をもとにしたデータ集合から，都道府県と市郡の包含関係を推定し，提案手法の有効性を確認した．

2. 関 連 研 究

条件付き確率を用いる古典的な問題として，相関ルールマイニングがある．相関ルールマイニングは，複数のアイテムからなるトランザクションの集合から関係の強いアイテムの組み合わせを発見する，データマイニングの主要技術である．アイテムの集合をアイテム集合といい，アイテム集合 X とアイテム集合 Y の間に成立する関係は相関ルール $X \Rightarrow Y$ として表される．ただし，$X \cap Y = \phi$ である．この関係は，あるトランザクションにアイテム集合 X が含まれるとき，そのトランザクションにアイテム集合 Y も含まれるという関係である．相関ルール $X \Rightarrow Y$ は強さをもち，その強さは条件付き確率 $P(Y|X)$ として表される．条件付き確率 $P(Y|X)$ の値が高いほど，アイテム集合 X とアイテム集合 Y が同じトランザクションに含まれる傾向にあることを意味している．相関ルールマイニングでは，相関ルールの強さを条件付き確率として推定し，その推定値をもとに関係の強いアイテムの組み合わせを発見する．そのため，条件付き確率の推定方法が相関ルールを発見する手法の性能に直接影響を与える．

データ集合から相関ルールを発見する代表的手法として Apriori が提案されている [4]．この手法は関係の強さを測る尺度として，支持率で表現される信頼度を用いる．信頼度は式 (1) のように定義される．

$$\hat{c}([X \Rightarrow Y]) = \frac{s(X \cup Y)}{s(X)} \tag{1}$$

ここで，支持率 $s(X)$ はデータベースの全トランザクションに対するアイテム集合 X を含むトランザクションの割合，支持率 $s(X \cup Y)$ はアイテム集合 X とアイテム集合 Y をともに含むトランザクションの割合である．信頼度はアイテム集合 X を含むトランザクションに対するアイテム集合 Y を含むトランザクションの割合である．言い換えると，トランザクションにアイテム集合 X が含まれるとき，アイテム集合 Y が含まれる条件付き確率 $P(Y|X)$ の最ゆう推定値となる．Apriori では式 (2) を満たす相関ルールの信頼度を計算する．

$$s(X \cup Y) \geq Minsup \tag{2}$$

Apriori では，最小支持率（Minsup）というしきい値を設けて，それ以上の支持率をもつ相関ルールの信頼度を求める．一般的に，この最小支持率はユーザが指定し，その値は関係推定の対象となるデータ集合によって異なる．最小支持率を設けることによって，統計的に不安定な相関ルールを無視して信頼度を計算できる．しかし，支持率が低く信頼度が高い相関ルールを推定できなくなるという問題がある．

一方で，低頻度の相関ルールを発見する研究も行われている．文献 [5] は，仮説検定によるスコアをもとに低頻度の相関ルールを発見する手法を提案した．文献 [6] は，配列に基づく枠組みとハッシュに基づく枠組みを提案した．これらの手法は，条件付き確率に基づいて相関ルールを発見する単純な手法ではないため，本論文で踏み込んだ議論はしない．

Apriori の問題に対処するため，ベイズの枠組みを導入した PredictiveApriori が提案されている [7]．この手法は関係の強さを測る尺度として，事後分布の期待値を用いる．条件付き確率 $P(Y|X)$ の真値は事後分布から予測される関係の正確さを表す値であり，そのときの関係の強さを $c([X \Rightarrow Y])$ とする．また，二項分布を $B[c,s](\hat{c})$ とする．このとき，$c([X \Rightarrow Y])$ の期待値はベイズの枠組みを用いて，式 (3) のように表される．

$$E(c([X \Rightarrow Y]) \mid \hat{c}([X \Rightarrow Y]), s(X)) = \frac{\int cB[c, s(X)](\hat{c}([X \Rightarrow Y]))\pi(c)dc}{\int B[c, s(X)](\hat{c}([X \Rightarrow Y]))\pi(c)dc} \tag{3}$$

$c([X \Rightarrow Y])$ の期待値は $\pi(c)$ を事前分布とした事後分

布の期待値である．この値を計算するためには積分を行う必要がある．そのため，関係の強さ c を任意の離散区間に分割し，各区間の中間点を加算することで積分の近似を行う．また，事前分布の取り扱いも事後分布の期待値を計算する過程で問題となる．真の事前分布を導出するためには，全ての相関ルールが存在しうる空間を調べなければならない．そこで，一様分布の仮定の下でランダムに多くの相関ルールをサンプリングする．そして，離散化した関係の強さ c の値を階級とする．各階級の値をもつ相関ルールの頻度についてヒストグラムを作成し，事前分布 $\pi(c)$ とする．

製品の品質をコントロールする必要があるとき，製品に不良品がある確率の信頼区間を構成する．そして，不良である割合を推定するために，信頼区間の上限値を用いる．ここで，信頼区間の上限値を用いる理由は，不良品を避けるためである．このときは，通常の製品が不良品であると判定されるよりも，不良品が通常の製品であると判定される方が被る損害が大きくなる場合がある．条件付き確率を用いて関係の強さを推定する場合は，良品にあたる真の関係を取りたいと考える．このときは，推定結果の適合率について注意する必要がある．適合率が 50%では，得られる結果が不十分であると考える．これは，真の関係を偽であると判定するよりも，偽の関係を真であると判定する方が大きな損害となることを意味している．この場合は，関係の強さを測る尺度として，信頼区間の下限値を用いることが妥当である．いま，事象の発生頻度が低いときを想定する．観測された製品のサンプル数が少ない場合，そのサンプル数をもとに構成される信頼区間の幅は広くなり，信頼区間の上限値は大きくなる．このことから，上限値を用いることで不良品が存在する確率を安全のため，高めに見積もることができる．一方で，関係の強さを推定する場合は信頼区間の下限値を用いる．関係の観測頻度が低いとき，その頻度をもとに構成される信頼区間の下限値は小さくなる．このことから，下限値を用いることで真の関係がある確率を安全のため，低めに見積もることができる．本論文では確率を低めに見積もることを保守的な推定と表現する．

この保守的な推定を行うためには，低頻度の関係について信頼区間を構成する必要がある．低頻度の事象から信頼区間を構成する場合は注意しなければならない．信頼区間を漸近的に近似する手法は多く存在する．しかし，これらの手法は信頼区間を構成したい事象について，十分な観測頻度があることを仮定している．観測頻度が十分ではない場合，構成した信頼区間に誤差が生じることが明らかになっている．製品から不良品を発見するために十分なサンプル数を観測し，信頼区間の近似公式を用いることができても，関係の強さを推定する場合に観測できるサンプル数には限りがある．また，信頼区間のいわゆる “正確な公式” [8] においても，観測頻度が不足している場合は信頼区間に誤差が生じることが明らかになっている [9]．以上のことから，本論文では，信頼区間を数値的に計算する．

3. 問題設定

あるデータ集合 $D = \{t_1, t_2, \cdots, t_n\}$ に存在するアイテム集合を $I = \{i_1, i_2, \cdots, i_m\}$ とする．データ集合を構成する各要素 $t_k (t_k \subseteq I)$ をトランザクションと呼び，長さ m のアイテム集合とは m 個のアイテムの組み合わせによって構成されたアイテム集合のことである．ここで，各アイテムを実世界に存在する地名（都道府県市郡）と考えるとアイテム集合 I の例は次のようになる．

$I = \{$ 東京都, 大阪府, 神戸市, 北海道, 江別市 $\}$

$\langle x, y \rangle$ は一つのアイテムからなるアイテム集合 x と y の間で定義される包含関係である．アイテム集合 x に含まれるアイテムは，別のアイテム集合 y に含まれるアイテムを概念的に包含する．この $\langle x, y \rangle$ は，次の二つの定義を満たす．

$$\forall a \subset I; \forall b \subset I; \forall c \subset I; \\ \langle a, c \rangle \wedge \langle b, c \rangle \rightarrow a = b \qquad (4)$$

$$\exists a \subset I; \exists b \subset I; \exists c \subset I; \\ \langle a, b \rangle \wedge \langle a, c \rangle \rightarrow b \neq c \qquad (5)$$

式 (4) は，関係の右のアイテム集合が等しい場合は左のアイテム集合も等しいという定義である．式 (5) は，この関係には一対一ではないアイテム集合が存在するという定義である．これらの定義を満たす包含関係 $\langle x, y \rangle$ の集合を R と定義し，正解集合と呼ぶことにする．本論文では，データ集合から正解集合 R を推定するという問題を扱い，条件付き確率を推定する手法の性能比較を行う．アイテム間の包含関係を推定対象とすることによって，条件付き確率が関係の強さを計ることに適した尺度となる．この条件付き確率は，包含するアイテムが出現するという条件下で包含されるアイテムが出現する確率となる．一つのアイテムからな

るアイテム集合をそれぞれ $S_c = \{i_c\}$，$S_p = \{i_p\}$ とすると，正解集合 R は次のように定義される．

$$R = \{\langle S_c, S_p\rangle \mid S_c, S_p \subset I\}$$

上記の定義を満たす関係の一つとして都道府県と市郡の包含関係が考えられる．S_c をある都道府県からなるアイテム集合，S_p をそこに属する市郡からなるアイテム集合とすると正解集合 R の例は次のようになる．

$$R = \{\langle \{\text{ 北海道 }\}, \{\text{ 札幌市 }\}\rangle, \langle \{\text{ 北海道 }\}, \{\text{ 釧路市 }\}\rangle\}$$

包含関係 $\langle S_c, S_p\rangle$ は相関ルールを用いて $S_c \Rightarrow S_p$ と表すことができる．例えば，正解集合 R の例に含まれる関係は相関ルールを用いて，{ 北海道 } ⇒ { 札幌市 }，{ 北海道 } ⇒ { 釧路市 } と表すことができる．

本論文では，次の手順で正解集合 R を推定する．まず，各トランザクションに含まれる二つのアイテムからなる全ての組を相関ルールとして求め，頻度情報をもとに相関ルールの強さである条件付き確率を推定する．そして，その推定値が高いほどアイテム間の関係性が高いと推定する．最後に，各トランザクションに含まれる二つのアイテムからなる全ての組がもつ関係 $\langle S_c, S_p\rangle$ が R に含まれるかどうかで包含関係の正解判定を行う．

4. 提案手法

事象の観測頻度から条件付き確率 $P(Y|X)$ の真値を推定するとき，一般的に最ゆう推定値や期待値が推定値として用いられる．本論文では，$P(Y|X)$ の推定値を小さめに見積もる方が安全と考え，保守的な推定を行う方法を提案する．提案手法は，事前分布として何らかの分布を仮定し，事後分布を求めて信頼区間の下限値を推定値とする．まずはくじ引きを例として提案手法の基本的アイディアについて述べる．そして，このアイディアをデータマイニングの問題に適用する一例として，相関ルールマイニングへの応用方法を述べる．

4.1 基本的アイディア

確率 θ で当たりが含まれるくじがあるとする．θ の値が 0 から 1 の範囲で等確率に決まると仮定すると，θ の事前分布 $\pi(\theta)$ は図 1 のような一様分布になる．このとき，θ の期待値 $\bar{\theta}$ は 1/2 となる．また，$\bar{\theta}$ に対する片側 $100(1-\alpha)$%信頼区間の下限値は α となり，これは $100(1-\alpha)$%の確率で $\bar{\theta}$ の値が α 以上となることを意味している．ここで，片側 $100(1-\alpha)$%は信頼係数と呼ばれ，信頼区間を求める際のパラメータとなる．事前分布 $\pi(\theta)$ を図 1 の一様分布と仮定すると，くじを n 回引いて x 回当たりを観測したときの事後分布 $P(\theta \mid N = n, X = x)$ は次のようになる．

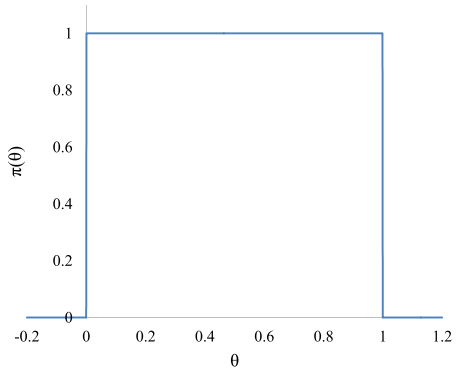


図 1　θ の一様分布
Fig. 1　Uniform Distribution for θ.

$$P(\Theta \mid N = n, X = x) = \begin{cases} L \times \theta^x (1-\theta)^{n-x} & (0 < \theta < 1), \\ 0 & \text{Otherwise.} \end{cases} \quad (6)$$

ただし，L は $P(\Theta \mid N = n, X = x)$ を $(-\infty, \infty)$ の範囲で積分したときに 1 とするような正規化定数である．θ の期待値 $\bar{\theta}$ は式 (7) に示すように，ラプラススムージングの値となることが知られている [3]．

$$\bar{\theta} = \int \Theta \cdot P(\Theta \mid N = n, X = x) d\Theta = \frac{x+1}{n+2} \quad (7)$$

$\bar{\theta}$ に対する片側 $100(1-\alpha)$%信頼区間 $[\theta_{lb}, 1]$ は次のようになる．

$$P(\Theta > \theta_{lb} \mid N, X) = 1 - \alpha \quad (8)$$

いま，くじを 1 回引いて 0 回当たりを観測した場合とくじを 4 回引いて 1 回当たりを観測した場合について考える．このとき，それぞれの事後分布 $P(\Theta \mid N = 1, X = 0)$ と $P(\Theta \mid N = 4, X = 1)$ は図 2，図 3 のようになる．ここで，$P(\Theta \mid N = 1, X = 0)$ において，信頼区間の下限値 θ_{lb} は 0 より大きくなる．この下限値 θ_{lb} は，観測しなかった事象に保守的に確率を割り当てた値となる．各グラフからわかるように，$\bar{\theta}$ は両方とも 1/3 となる．しかし，$P(\Theta \mid N = 4, X = 1)$ は

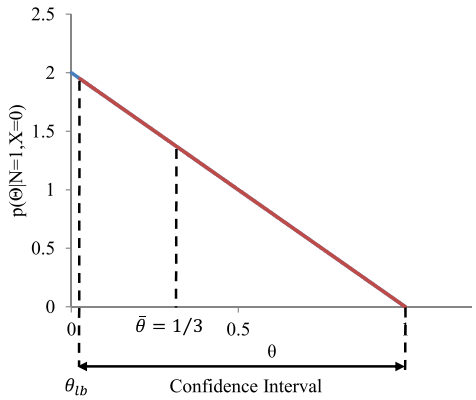


図 2　事後分布 $P(\Theta \mid N=1, X=0)$ と信頼区間 $[\theta_{lb}, 1]$
Fig. 2　Posterior Distribution $P(\Theta \mid N=1, X=0)$, and its confidence interval $[\theta_{lb}, 1]$.

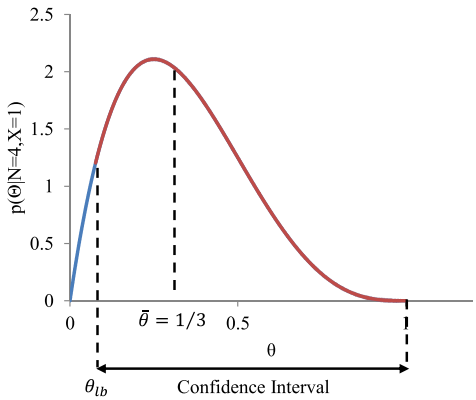


図 3　事後分布 $P(\Theta \mid N=4, X=1)$ と信頼区間 $[\theta_{lb}, 1]$
Fig. 3　Posterior Distribution $P(\Theta \mid N=4, X=1)$, and its confidence interval $[\theta_{lb}, 1]$.

$P(\Theta \mid N=1, X=0)$ よりも分散が小さいため，θ_{lb} が大きくなる．信頼区間の下限値を用いると，事後分布の分散を考慮し，θ の推定値を小さめに見積もることができる．この推定値は，確率を保守的に推定した値となる．

4.2　相関ルールマイニングへの応用

条件付き確率 $P(Y|X)$ の真の値を θ とする．ここで，データベースにあるアイテム集合 X を含むトランザクション数を n，アイテム集合 X と Y を含むトランザクション数を k とする．これらは，**4.1** のくじを引いた回数と当たりを観測した回数にそれぞれ対応する．いま，データベースのトランザクション数を $|D|$ とすると，支持率 $s(X)$ と $s(X \cup Y)$ はそれぞれ $n/|D|$，$k/|D|$ とみなすことができる．このとき，信頼度は **2.** の式 (1) で示したように，支持率を用いて $s(X \cup Y)/s(X)$ と定義され，$P(Y|X) = \theta$ の最ゆう推定値 $\hat{\theta} = k/n$ と解釈することもできる．

事前分布 $\pi(\theta)$ として何らかの分布を仮定し，θ の事後分布を求め，θ の期待値 $\bar{\theta}$ に対する片側 $100(1-\alpha)$%信頼区間を求める．本論文では，この信頼区間の下限値を $P(Y|X)$ の推定値とする．信頼区間の計算については，近似公式による誤差が影響すると考え，直接に数値積分で求める [9]．

信頼度と支持率はトレードオフであり，どちらを重要視するかはアプリケーションによって異なる．最小支持率を高くして支持率を重要視すると，出現頻度は高いが信頼度の低い相関ルールが発見される．逆に，最小支持率を低くして信頼度を重要視すると，出現頻度は低いが信頼度の高い相関ルールが発見される．本研究では，信頼度が同じでも支持率の値によって相関ルールの価値は異なるという考え方を実現する．この考え方に基づくと，支持率が低い場合に相関ルールの価値も相対的に低くなることが望ましい．信頼区間の下限値を $P(Y|X)$ の推定値とすることによって，低頻度で高い信頼度をもつ相関ルールの強さは弱くなる．

5.　実　　験

提案手法の有効性を確認するために，提案手法と Apriori，θ の期待値（ラプラススムージングの値）を推定値として用いた場合，PredictiveApriori について条件付き確率に基づく関係の推定性能を比較する．提案手法は事前分布として何らかの分布を仮定する必要がある．ここでは，必然性のあるものは特定しにくいこと，及び，θ の期待値との比較を明瞭にするという理由から，事前分布として一様分布を仮定する．関係の推定対象としては，文献 [10], [11] と同様に地名（都道府県市郡名）を用いた．地名を用いる理由は，実世界において地名間には包含関係が成り立っていること，地名間の包含関係は定まっているため，正解判定が容易であること，そして関係の強さを測る尺度として条件付き確率が適切であることが挙げられる．実験では，人工的に生成したデータ集合と実際の新聞記事をもとにしたデータ集合を用い，関係の強さを測る各手法のふるまいを観測する．

5.1　実験で用いるデータ集合

実験では，二種類のデータ集合を用いた．一つ目は実世界の地名間にある包含関係から二種類の組を取り出し，それらの組に現れる地名からなるトランザクションを要素とする，人工的に生成したデータ集合で

Algorithm 1 データ集合の生成アルゴリズム

```
D* := φ; k := 0;
while k < 1000 do
  j := 0; t_k := φ;
  while j < 2 do
    R からランダムに ⟨S_c, S_p⟩ を取り出す;
    t_k := t_k ∪ S_c ∪ S_p;
    j := j + 1
  end while
  D* := D* ∪ {t_k};
  k := k + 1
end while
```

ある．二つ目は実際の新聞記事コーパスから得られるデータ集合である．これらのデータ集合について以下で述べる．

- 人工データ：実データを用いた実験では，アイテム集合に対する出現頻度の偏りが観測される．この偏りを考慮せずに，提案手法における包含関係の推定性能を評価するため，人工的に生成した四つのデータ集合を用いた．これらのデータ集合は文献 [10] と同様の方法で正解集合 R から生成された．この生成アルゴリズムをアルゴリズム 1 に示す．

生成したデータ集合に含まれるトランザクションは，例えば，$t_k = \{$ 北海道, 札幌市, 愛知県, 名古屋市 $\}$ などが考えられる．これは正解集合 R から $\langle\{$ 北海道 $\}, \{$ 札幌市 $\}\rangle$ と $\langle\{$ 愛知県 $\}, \{$ 名古屋市 $\}\rangle$ という二種類の組の関係を抽出して，これらの関係に現れる四つのアイテムからなるトランザクションである．データ集合のトランザクションの数を無限に増やすことはできないため，このようなトランザクションを 1,000 個もつデータ集合を生成した．これらのデータ集合に関する情報を表 1 に示す．正解集合 R に含まれる全正解数は 1,215 であるが，データ集合に含まれない正解があることによって，実態に合った評価ができると考える．生成したデータ集合を用いて，トランザクションごとに組み合わされた二種類の組の関係を各手法の値によって正しく分離することを試みる．

- 実データ：実世界では，アイテムの出現頻度に偏りがあるデータ集合から関係を推定する必要がある．そこで，毎日新聞記事コーパス（91 年～97 年版）をもとにしたデータ集合を用いた．これらのデータ集合は，一記事を一つのトランザクション，記事に出現する地名（都道府県市郡名）をそのトランザクションに含まれるアイテムとする．これらのデータ集合に関する情報を表 2 に示す．

実データには「静岡」などのように都道府県市郡を表す語が付いていないアイテムも含まれる．このようなアイテムは都道府県市郡を表す語を付与し，実世界に存在する関係が成り立てば正解とみなす．例えば，「静岡」について「静岡県」とすると正解集合 R に含まれる場合,「静岡市」とすると正解集合 R に含まれる場合のいずれかであれば正解とみなす．新聞記事コーパスの各年版に含まれる正解数が正解集合 R の全正解数 1,215 よりも多いのは，都道府県市郡を表す語が付いていないアイテムを含む関係も正解として数えたためである．新聞記事コーパスをもとにしたデータ集合を用いて，新聞記事に現れる地名の組から実世界に存在する地名の組を推定する性能を測定する．

表 1　人工データの特性
Table 1　Properties of the synthetic datasets.

	データ集合			
	1	2	3	4
トランザクション数	1,000	1,000	1,000	1,000
候補となる組の種類	4,469	4,499	4,438	4,453
候補となる組の出現数	5,934	5,490	5,913	5,901
正解集合に含まれる組の種類	975	993	979	984
正解集合に含まれる組の出現数	2,000	2,001	2,001	2,001

表 2　実データの特性
Table 2　Properties of the real-world datasets.

	毎日新聞コーパスの年版						
	91	92	93	94	95	96	97
トランザクション数	52,232	56,587	52,031	65,922	76,563	58,537	71,955
候補となる組の種類	252,139	223,927	210,927	253,712	226,933	162,819	161,662
候補となる組の出現数	681,010	786,917	782,845	1,034,728	1,271,310	658,548	767,057
正解と判断される組の種類	3,790	3,654	3,660	3,960	3,706	3,536	3,445
正解と判断される組の出現数	63,142	65,045	60,699	74,833	103,884	61,177	73,099

5.2 評価手順

実験の評価手順は次に示すとおりである．まず，各トランザクションに含まれる二つの地名からなる組を全て求め，評価対象となる各手法の値を計算する．そして，手法の値が高いほど組の関係性が高いと判断し，その値によって組を降順に並べてランク付けをする．最後に，ランクが上位となる関係から順に正誤を確認し，横軸をランク，縦軸を再現率として上位からランク—再現率曲線を描く．人工データによる実験では上位 4,000 件，実データによる実験では上位 12,000 件までの関係を正誤確認の対象とする．再現率の定義を次に示す．

$$\text{再現率} = \frac{\text{あるランクまでの正解数}}{\text{データ集合に含まれる正解数}}$$

ランク—再現率曲線によるランクの上位に着目し，各手法について，都道府県市郡の包含関係を推定する性能を再現率と適合率の観点から評価する．

5.3 パラメータ設定

実験で用いる手法は，それぞれ固有のパラメータをもつ．そのため，最適なパラメータを設定し，最良の性能を比較することで手法間の公平な性能評価を行う．ここで述べる最良の性能とは，それぞれの手法が最も効率良く正しい関係を推定できる性能のことである．

Apriori は最小支持率というパラメータをもつ．人工データを用いた実験では，最小支持率を変化させてその振る舞いを観察する．実データを用いた実験では，ランクの上位で適合率が高く，下位で再現率を保持するような最小支持率を探し，$5/|D|$ という値を設定した．

提案手法は信頼係数というパラメータをもつ．人工データを用いた実験，実データを用いた実験において，ランクの上位で適合率が高く，下位で再現率を保持するような信頼係数を探した．そして，両実験ともに片側 99% ($\alpha = 0.01$) という値を設定した．

PredictiveApriori は事後分布を計算するために，事前分布を推定する必要がある．まず，データ集合から同じトランザクションに含まれる二つの地名をランダムに取り出す．この試行を繰り返し，人工データでは存在する組を全て取り出す．実データでは 10 万個の組を取り出す．そして，取り出した組について θ のヒストグラムを作成する．人工データから作成したヒストグラムはどのデータ集合においても同様の形状となる．本論文では，掲載できる紙面の都合により，データ集合 1 のヒストグラムのみを図 4 に示す．なお，その他のデータ集合から作成したヒストグラムはウェブサイト(注1)にて公開する．実データから作成したヒストグラムはどの年版においても同様の形状となるため，91 年版のみを図 5 に示す．

ここで，人工データのヒストグラムに着目する．図 4 のグラフから，θ の値が 1/1，1/2，1/3 といった特定の値でヒストグラムが急上昇していることがわかる．このような θ の値をもつ組はいずれも出現頻度が低い．この急上昇の原因として，例えば θ の真値が 0.4 などであった場合に，ヒストグラム上では θ の値が 1/2 といった特定の値に吸い込まれてしまうことが考えられる．これは，ヒストグラムによって θ の分布を離散化した影響によるものである．PredictiveApriori はデー

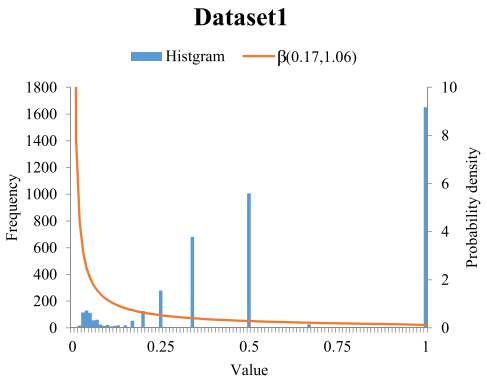


図 4 人工データにおけるヒストグラムとベータ分布
Fig. 4 Histgram and beta distribution for the synthetic dataset.

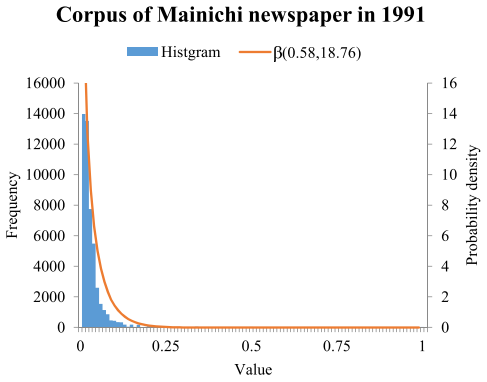


図 5 実データにおけるヒストグラムとベータ分布
Fig. 5 Histgram and beta distribution for the real-world dataset.

(注1)：http://www.ss.cs.tut.ac.jp/IEICE29-04_results/

タ集合から得られたヒストグラムを事前分布とする．しかし，人工データのヒストグラムについて，これを事前分布として仮定することは適切ではない．関係に成り立つ強さの真値 θ は 0 から 1 までの範囲で値を取るが，1/2 などの特定の値である理由はない．

図 4 において，推定すべきヒストグラムのピークは 0.04 付近にあると考えられる．そこで，このピークよりも大きいヒストグラムを取り除くことにする．具体的には，人工データのヒストグラムから 0，1/1，1/2，1/3，2/3，1/4，3/4，1/5，2/5，3/5，4/5 を取り除き，θ の平均と分散を求める．そして，平均と分散が一致するようにベータ分布を推定する．ベータ分布は次のように定義される．

$$\beta(a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 t^{a-1}(1-t)^{1-b}dt} \tag{9}$$

ベータ分布は二つのパラメータ a と b をもつ．これらのパラメータを調節することで，ベータ分布を推定する．データ集合 1 から推定したベータ分布 $\beta(0.17, 1.06)$ を図 4 に示す．このベータ分布は θ の値が小さいときに確率密度が大きくなるような形状になる．他のデータ集合においても，ヒストグラムから 0〜4/5 を取り除いてベータ分布を推定した．これらのベータ分布はウェブサイト(注1)にて公開する．実験では，推定したベータ分布を人工データにおける事前分布とする．ベータ分布を事前分布とすることによって，推定するパラメータ数を二つに減らすことができる．

実データのヒストグラムでは，人工データで観測されるようなある値でのヒストグラムの急上昇はない．そのため，ヒストグラムをそのままベータ分布の推定に利用することができる．θ の平均と分散を求め，それらが一致するようにベータ分布を推定する．91 年版のベータ分布を図 5 に示す．ベータ分布は θ の値が小さいときに確率密度が大きくなるような形状になる．ヒストグラムの形状を見ると同様の傾向を示しており，ベータ分布はヒストグラムを適切に近似できているといえる．他の年版でもベータ分布を用いることにより，ヒストグラムを適切に近似できることを確認した．実験では，推定したベータ分布を実データにおける事前分布とする．

5.4 人工データによる実験結果

人工データによる実験結果はどのデータ集合においても同様の傾向を示す．本論文では，掲載できる紙面の都合により，データ集合 1 による結果のみを図 6〜図 8 に示す．なお，その他のデータ集合による結果は **5.3** の脚注で示したウェブサイト(注1)にて公開する．これらのグラフは，横軸をランク，縦軸を再現率とするランク—再現率曲線である．グラフ上の点と原点

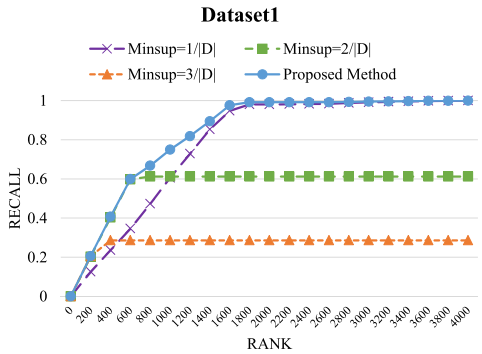


図 6　人工データにおける Apriori との比較
Fig. 6　A comparison with Apriori for the synthetic dataset.

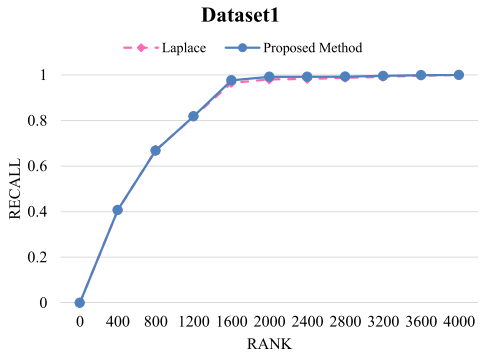


図 7　人工データにおける θ の期待値との比較
Fig. 7　A comparison with the expected value of θ for the synthetic dataset.

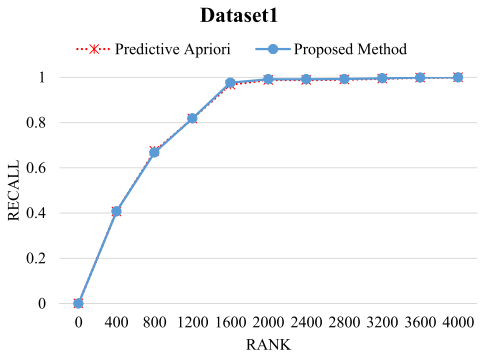


図 8　人工データにおける PredictiveApriori との比較
Fig. 8　A comparison with PredictiveApriori for the synthetic dataset.

を結んだ線の傾きが適合率に比例する．同一のランクにおいて，曲線が上にある手法ほど，そのランクにおいて優れた性能をもつ．各図はそれぞれ，提案手法と Apriori との比較結果，θ の期待値との比較結果，PredictiveApriori との比較結果を示す．$|D|$ はデータ集合に含まれるトランザクション数である．最小支持率が $1/|D|$ のときは，データ集合に 1 回以上含まれる組を関係推定の対象とするため，最小支持率を設けないことを意味する．

図 6 に示すように，最小支持率を $1/|D|$ として，データ集合に含まれる全ての組について信頼度を計算すると，頻度が低いにもかかわらず，信頼度が高い組（不正解となる組）が推定値のランク上位となる傾向がある．この傾向は上位の適合率を低下させる原因になる．そこで，最小支持率を $2/|D|$ にすると，出現頻度が 2 以上の組が上位となり，そのような不正な組が取り除かれるため，上位の適合率は向上したと考えられる．しかしながら，データ集合において頻度が 1 となる組は一律に推定対象から取り除かれ，このときに正解の組も共に取り除かれてしまう．このため，下位の再現率が低下したと考えられる．

最小支持率を $2/|D|$ から $3/|D|$ に上げると上位の適合率は向上せず，下位の再現率が更に低下した．これは最小支持率を上げすぎたため，正解と判断される多くの組が取り除かれてしまうことが原因と考えられる．このことから，最小支持率はこれ以上変化させても意味がなく，$2/|D|$ を最適な最小支持率とすることが妥当と考えられる．

提案手法は，ランク上位では，最適な最小支持率を設けた信頼度とほぼ等しい適合率となる．このことから，提案手法は最適な最小支持率の信頼度と同様に，不正解の組を取り除いていると考えられる．下位では，最適な最小支持率の信頼度よりも高い再現率となる．これは頻度の低い組を正しく扱っていると考えられる．

図 7 からわかるように，θ の期待値は提案手法とほぼ同じ性能を示す．ここで，**5.3** で示した図 4 からわかるように，人工データには θ の値が 1 となるような関係が多く存在する．このような関係には，頻度が低い不正解の組が多く含まれると考えられる．ラプラススムージングはこのような組の推定値を低く見積もり，上位から取り除くことができる．その結果として，θ の期待値は，都道府県と市郡の包含関係を効率良く推定できると考えられる．

PredictiveApriori は，図 4 に示すベータ分布を事前分布として用いた．この事前分布は，θ の値が小さいときに確率密度が大きくなる．地名の観測頻度が低いとき，事後分布の形状は事前分布に近づく．そのため，事後分布の期待値は信頼度よりも小さい値となり，提案手法と同様に関係の強さを保守的に見積もることになっている．図 8 からわかるように，PredictiveAprirori は提案手法とほぼ同じ性能を示す．実験では，事前分布としてベータ分布を用いることによって，ヒストグラムを用いるよりもパラメータ数を削減することができた．それでも，ベータ分布を用いるには二つのパラメータを推定しなければならない．一方で，提案手法は信頼係数というただ一つのパラメータを推定するのみでよいという特徴がある．

5.5 実データによる実験結果

実データによる実験結果はどの年版においても同様の傾向を示す．そのため，91 年版の結果のみを図 9〜図 11 に示す．各グラフの見方は人工データによる実験結果と同様である．各図はそれぞれ，提案手法と Apriori との比較結果，θ の期待値との比較結果，PredictiveApriori との比較結果を示す．

図 9 に示すように，最小支持率を $1/|D|$ とすると，人工データを用いた実験と同様に適合率が低下する傾向が見て取れる．実データを用いた実験では，人工データを用いた実験よりもこの傾向が強く表れ，上位の適合率を大きく低下させる原因になる．

そこで，人工データによる実験と同様に，最小支持率を $1/|D|$, $2/|D|$, $\cdots$, $6/|D|$ と変化させて実験を行い，$5/|D|$ という値を最適な最小支持率とみなした．これによって，上位における適合率は向上したが，下

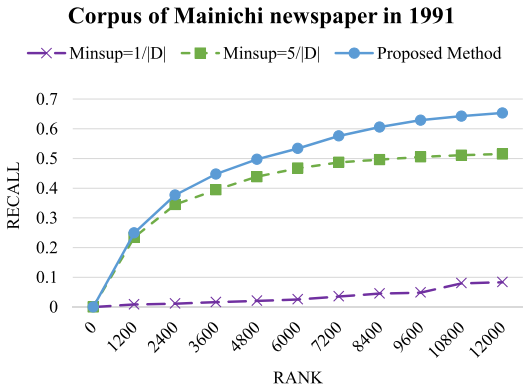


図 9 実データにおける Apriori との比較
Fig. 9 A comparison with Apriori for the real-world dataset.

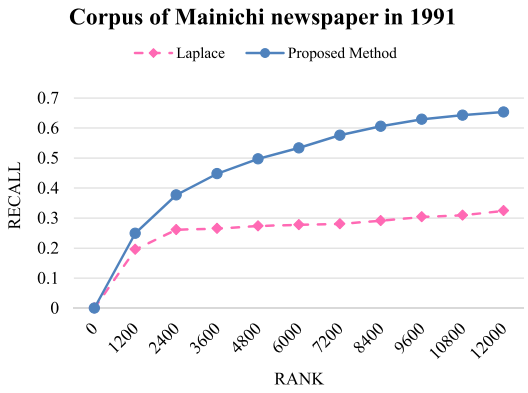


図 10　実データにおける θ の期待値との比較
Fig. 10　A comparison with the expected value of θ for the real-world dataset.

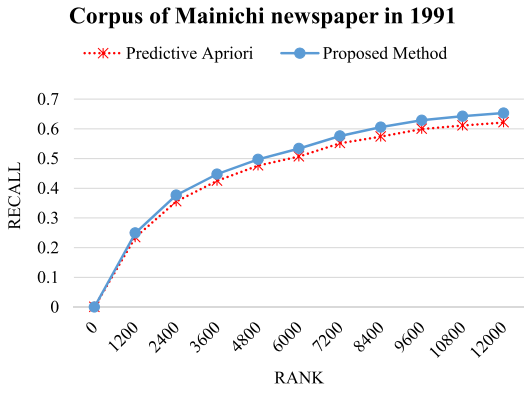


図 11　実データにおける PredictiveApriori との比較
Fig. 11　A comparison with PredictiveApriori for the real-world dataset.

位の再現率は低下した．なお，最適な最小支持率は人工データによる実験では $2/|D|$ であったが，実データによる実験では $5/|D|$ となった．これは，人工データと比較して実データのサイズが大きく，それに伴って含まれる不正解となる組も多くなったためと考えられる．

提案手法は，ランク上位では最適な最小支持率の信頼度とほぼ等しい適合率となり，下位では最適な最小支持率の信頼度よりも高い再現率となった．以上のことから，提案手法は，アイテムの出現頻度に偏りがある実データに適用した場合においても，低頻度を適切に対処できることを示唆した．

図 10 からわかるように，θ の期待値は上位の適合率が提案手法とほぼ同じとなる．これは，人工データを用いた実験と同様に，頻度が低いにもかかわらず，θ の値が高い組が取り除かれることが原因と考えられる．一方で，下位の再現率は提案手法よりも低い．ラプラススムージングは θ の値が 1/2 よりも低い組，すなわち，あまり共起しない組の推定値を高く見積もってしまう．この作用によって，θ の期待値は，下位において再現率の低下を招いたと考えられる．

PredictiveApriori は，図 5 に示すベータ分布を事前分布として用いた．PredictiveApriori は，事前分布の働きによって関係の強さを保守的に見積もることになっている．図 11 からわかるように，PredictiveApriori は提案手法とほぼ同じ性能を示すが，提案手法は推定すべきパラメータ数が一つでよいという特徴がある．

6. 考　　察

条件付き確率は，データマイニングの分野でよく用いられ，その推定方法によって手法の正確さが左右されることがある．本論文では，事前分布を一様分布として事後分布を求め，信頼区間の下限値を条件付き確率の推定値とすることを提案した．そして，相関ルールマイニングの観点から条件付き確率の推定という問題に取り組み，提案手法の有効性を確認した．

Apriori は関係の強さを測る尺度として信頼度を用いる．信頼度は条件付き確率の最ゆう推定値である．しかし，信頼度（最ゆう推定値）は低頻度の影響を受けやすい．そこで，Apriori は最小支持率というしきい値を設け，出現頻度がそれに満たない関係は推定の対象としない．最小支持率を設けることで，効率的に関係を推定できるが，このときに低頻度で高い信頼度をもつ関係も推定対象から取り除いてしまう．このとき，取り除かれた関係の中には，都道府県と市郡の正しい包含関係が多く含まれることが実験から明らかになった．これらの関係は発見されるべきである．実験では，提案手法が最小支持率を変化させたどの信頼度よりも高い性能を示した．この結果は，提案手法が低頻度の確率推定に有効であることを意味している．

条件付き確率を保守的に推定することの有効性を確認するため，θ の期待値を推定値として用いた場合との比較実験を行った．提案手法が条件付き確率を保守的に推定する一方で，θ の期待値は偏りのない推定値となる．θ の期待値は **4.1** の式 (7) に示したようにラプラススムージングの値であるため，解析的に求めることができる．一方で，提案手法は信頼区間を数値積分で求める必要があり，θ の期待値よりも要する計算量が多い．実験では，θ の期待値はランク上位において提案手法とほぼ同じ適合率となった．そのため，ラ

ンク上位で関係を発見する上では，少ない計算量で導出できる θ の期待値が優れている．しかし，ラプラススムージングの値は偏りのない推定値であるため，あまり共起しない関係の強さに正のバイアスをかけてしまう．実データを用いた実験では，θ の期待値を用いると，前述の作用によって下位の再現率が低下することが示唆された．このことから，下位でも効率良く関係を発見する場合には，条件付き確率を保守的に推定する提案手法が有効と考えられる．

本論文では，提案手法と PredictiveApriori の性能比較も行い，両者がほぼ同じ性能を獲得しうることを示唆した．しかし，データをもとに事前分布を推定するという PredictiveApriori の考え方と，結果の利用するときの適合率で関係の推定値を調整するという提案手法の考え方は異なる．一般的に，データをもとに事前分布を推定することは容易ではない．PredictiveApriori では事前分布を推定するために，データ集合からヒストグラムを作成する必要がある．人工データを用いた実験では，**5.3** で述べたようにヒストグラムから θ の値が 0，1/1，1/2，1/3，2/3，1/4，3/4，1/5，2/5，3/5，4/5 となる低頻度の組を取り除くことで事前分布を推定した．しかし，これが本当に正解情報を使用していないのかということについては疑念が残る．実験では，事前分布としてヒストグラムの代わりにベータ分布を用いた．これによって推定すべきパラメータ数を削減したが，それでも二つのパラメータを推定する必要がある．それに対し，提案手法は事前分布として一様分布を仮定したため，推定すべきパラメータは信頼係数のみとなる．以上のことから，この比較実験では，提案手法の方が概念的に単純で適用範囲が広いと考えられる．

一方で，事前分布として何を選択するかはよく議論される問題である [12]．当然，データについての事前分布を推定した上で，推定値を保守的に推定するということも考えられる．しかし，現状では，データをもとに事前分布を推定することは難しい場合が多く，事後分布から推定値を保守的に求める方法も不明である．そのため，このことは更に検討が必要である．本論文では，提案手法で用いる事前分布として一様分布を仮定したが，無情報事前分布としては Jeffreys 事前分布がよく用いられる [13]．そこで，データについての事前分布が推定困難な場合に，Jeffreys 事前分布を仮定し，推定値を保守的に推定することも今後の課題として挙げられる．

7. むすび

本論文では，条件付き確率の推定という問題に取り組み，概念的に単純で低頻度に対しても頑強な手法を提案した．提案手法は，事前分布として一様分布を仮定し，事後分布を計算する．そして，事後分布の信頼区間を求め，その下限値を条件付き確率の推定値とする．この推定値は条件付き確率を保守的に推定した値となる．提案手法について，Apriori との比較実験を行い，提案手法は低頻度の確率推定においても高い性能を示した．また，不偏な推定値である期待値を用いた場合との比較実験において，提案手法は下位で高い再現率を示し，推定値を保守的に見積もることの有効性を確認した．更に，ベイズの枠組みを導入した PredictiveApriori との比較実験も行い，両手法がほぼ同じ性能を獲得しうることを示した．だが，PredictiveApriori のように事前分布をデータに基づいて推定することは一般的に容易ではない．一方で，結果を利用するときの適合率に応じて推定値を調節する提案手法の考え方は直観的でわかりやすく，広い分野での応用が期待できる．

今後の課題として，データについての事前分布を推定した上で，条件付き確率を保守的に推定することが挙げられる．事前分布としてより適切な仮定を置くことによって，提案手法の性能向上を目指す．また，事前情報が得られない場合において，事前分布を Jeffreys 事前分布と仮定し，保守的な推定を行うことも検討したい．更に，提案手法のパラメータである信頼係数を自動的に決定する方法も検討する必要がある．

謝辞　本研究は，平成 27 年度岐阜聖徳学園大学研究助成金を受けた．

文　献

[1] T. Sonoda and T. Miura, “Conditional collocation in Japanese,” Proc. 18th Australasian Document Computing Symposium, pp.82–88, 2013.

[2] A. Jimeno-Yepes and R.B. Llavori, “Knowledge based word-concept model estimation and refinement for biomedical text mining,” J. Biomedical Informatics, vol.53, pp.300–307, 2015.

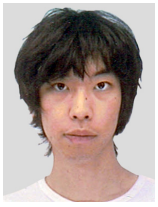
[3] 間瀬　茂，ベイズ法の基礎と応用：条件付き分布による統計モデリングと MCMC 法を用いたデータ解析，第 1 版，日本評論社，2016.

[4] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” Proc. 20th International Conference on Very Large Data Bases, pp.487–499, 1994.

[5] W. DuMouchel and D. Pregibon, “Empirical bayes screening for multi-item associations,” Proc. 7th

ACM SIGKDD international conference on Knowledge discovery and data mining, pp.67–76, 2001.
[6] L. Zhou and S. Yau, “Association rule and quantitative association rule mining among infrequent items,” Proc. 8th international workshop on Multimedia data mining:(associated with the ACM SIGKDD 2007), pp.1–9, 2007.
[7] T. Scheffer, “Finding association rules that trade support optimally against confidence,” Intelligent Data Analysis, vol.9, no.4, pp.381–395, 2005.
[8] C.J. Clopper and E.S. Pearson, “The use of confidence or fiducial limits illustrated in the case of the binomial,” Biometrika, vol.26, no.4, pp.404–413, 1934.
[9] M. Kikuchi, M. Yoshida, M. Okabe, and K. Umemura, “Confidence interval of probability estimator of Laplace smoothing,” Proceedings of the 2015 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA2015), pp.1–6, 2015.
[10] 山本英子，梅村恭司，“コーパス中の一対多関係を推定する問題における類似尺度，” 自然言語処理，vol.9, no.2, pp.45–75, 2002.
[11] 岡部正幸，梅村恭司，“頻度差が著しい場合における一対多関係を推定する類似尺度，” 情報学シンポジウム講演論文集，vol.2005, pp.129–136, 2005.
[12] A.R. Syversveen, “Noninformative bayesian priors. interpretation and problems with construction and applications,” Preprint Statistics, vol.3, pp.1–11, 1998.
[13] H. Jeffreys, “An invariant form for the prior probability in estimation problems,” Proc. Royal Society of London A: Mathematical, Physical and Engineering Science, vol.186, pp.453–461, 1946.

（平成 28 年 7 月 1 日受付，10 月 19 日再受付，
29 年 1 月 6 日早期公開）



菊地　真人

2016 豊橋技術科学大学情報・知能工学課程卒業．現在，同大学院博士前期課程在籍．主として，データマイニングに関する研究に従事．

山本　英子

1998 豊橋技術科学大学大学院工学研究科情報工学専攻修士課程修了．2002 同大学大学院工学研究科電子・情報工学専攻博士後期課程修了．博士（工学）．同年より独立行政法人情報通信研究機構有期研究員．2007 神戸大学工学研究科プロジェクト奨励研究員を経て，2009 神戸大学工学研究科講師．現在，岐阜聖徳学園大学経済情報学部准教授．自然言語処理に関する研究に従事．言語処理学会会員．

吉田　光男

2009 筑波大学第三学群情報学類卒業．2011 同大学院システム情報工学研究科博士前期課程修了，2014 同博士後期課程修了．博士（工学）．同年より豊橋技術科学大学大学院工学研究科（情報・知能工学系）助教．ウェブ工学，自然言語処理，計算社会科学に関する研究に従事．情報処理学会，言語処理学会，人工知能学会，日本データベース学会各会員．

岡部　正幸

1996 創価大学工学部情報システム学科卒業．2001 東京工業大学大学院総合理工学研究科知能システム科学専攻博士課程修了．博士（工学）．同年より科学技術振興機構 (CREST) 研究員．2003 豊橋技術科学大学情報メディア基盤センター助手，2007 同助教．2016 県立広島大学経営情報学部経営情報学科講師，現在に至る．情報検索，データマイニングに関する研究に従事．人工知能学会，ACM，IEEE-CS 各会員．

梅村　恭司　（正員）

1983 東京大学大学院工学系研究科情報工学専攻修士課程修了．博士（工学）．同年日本電信電話公社電気通信研究所入所．1995 豊橋技術科学大学工学部情報工学系助教授，2003 同教授．自然言語処理，システムプログラム，記号処理に関する研究に従事．情報処理学会，IEEE，電子情報通信学会，日本ソフトウェア科学会，言語処理学会，計量国語学会，ACM 各会員．