観測頻度に基づくゆう度比の保守的な直接推定

菊地　真人[†a)]　川上　賢十[†]　吉田　光男[†]　梅村　恭司[†]

Conservative Direct Estimation for Likelihood Ratios Based on Observed Frequencies

Masato KIKUCHI[†a)], Kento KAWAKAMI[†], Mitsuo YOSHIDA[†], and Kyoji UMEMURA[†]

あらまし　データを確率的に取り扱う問題において，統計的尺度の推定は手法の構成やデータ分析の基盤的役割を担う．本論文では統計的尺度の一つであるゆう度比を，離散的な標本空間から得た観測頻度をもとに推定する問題を扱う．素朴な推定方法は，ゆう度比の定義に従い，ゆう度比を構成する二つの確率分布を最ゆう推定して，その比を取ることである．しかし，低頻度からゆう度比を求めるとき，この方法は推定量を不当に高く見積もってしまう場合がある．そこで，ゆう度比の直接推定法 uLSIF を応用し，ゆう度比を低めに（保守的に）推定する方法を提案する．提案手法は，最ゆう推定によって求めたゆう度比を正則化パラメータによって調整する枠組みである．実験では提案手法の振る舞いを明らかにし，その有効性を示した．更に，自然言語処理におけるブートストラップ法を利用した実験も行い，提案手法の実用性も示した．

キーワード　ゆう度比，uLSIF，正則化，ブートストラップ法，保守的な推定

1. まえがき

ゆう度比は統計検定や多値分類 [1] などに多く用いられる尺度であり，ゆう度関数の比で表される．ゆう度比の真値を得ることは困難であるため，実際にゆう度比を用いる場合は推定が必要となる．自然言語処理やデータマイニングでは，観測頻度や標本に基づき，離散的な標本空間からゆう度比を推定する場合がある．素朴な推定方法は，ゆう度比の定義に従い，ゆう度比を構成する二つの確率分布を最ゆう推定して，その比を取ることである．しかし，この方法で低頻度からゆう度比を求めると，推定量を不当に高く見積もってしまう場合がある．

推定量を高く推定すると問題になる場合を考える．例えば，新聞記事コーパスに出現する地名から，後ろにカタカナ語が続きやすい地名を予測することを考える．コーパス中の文字列「東京タワー」を例にすると，「東京」が地名，「タワー」が後続するカタカナ語となる．「東京」と「豊橋」という地名が与えられたとき，次のゆう度比でカタカナ語の続きやすさを測るとする．

$$\mathrm{LR}_{\mathrm{MLE}}(a) = \frac{\hat{p}_1}{\hat{p}_2}$$
$$\hat{p}_t = \frac{k_t}{n_t} \qquad (t = 1, 2)$$

k_1 はある地名 a がカタカナ語直前，k_2 は a がコーパス全体に出現する頻度を表す．n_1 は全地名がカタカナ語直前，n_2 は全地名がコーパス全体に出現する総頻度を表す．このとき，$\hat{p}_1$ は a がカタカナ語直前，$\hat{p}_2$ は a がコーパス全体に現れる確率である(注1)．このゆう度比が 1 より大きいと a はカタカナ語直前に現れやすく，1 より小さいとコーパス全体に現れやすいことを意味する．コーパスから表 1 の頻度が観測できたとする．「東京」は k_2 が高く，コーパス全体で現れやすい．そして，カタカナ語直前での出現頻度 k_1 も高い．「豊橋」は「東京」よりも k_2 が低いため，相対的に k_1 も低くなる．特に，「豊橋」はカタカナ語の直前で 1 回しか現れず，推定値 $\mathrm{LR}_{\mathrm{MLE}}$(豊橋) の信頼性は $\mathrm{LR}_{\mathrm{MLE}}$(東京) よりも低いと考えられる．ところが，二

† 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
a) E-mail: m143313@edu.tut.ac.jp
DOI:10.14923/transinfj.2018DEP0007

(注1)：厳密には，コーパス中での単語（あるいは文字バイグラムなど）の出現を多項分布でモデリングしたとき，出現確率の最ゆう推定量を指す．

表 1　出現頻度の例
Table 1　The example of occurrence frequencies.

地名 a	k_1	k_2	n_1	n_2	$\mathrm{LR}_{\mathrm{MLE}}(a)$
東京	100	3,000	1,000	100,000	3.33
豊橋	1	30	1,000	100,000	3.33

つの地名について推定値を測ると両方とも 3.33 となってしまう．そこで，観測頻度が低い場合はゆう度比の正確な推定が難しいと考え，分母の確率推定量 $\hat{p}_2$ に応じて推定量を低めに見積もる方法を提案する．本論文では，推定量を低めに見積もることを “保守的な推定” と呼ぶ．

保守的な推定を実現するため，ゆう度比の直接推定法 uLSIF (unconstrained Least-Squares Importance Fitting) [2] に着目する．この方法は確率密度推定を介さずに，連続的なゆう度比の分布を直接推定する．具体的には，真のゆう度比と推定モデルとの誤差関数を定義し，その誤差を最小とする最適化問題を解く．一般にこの方法は推定量を保守的に見積もるものとされていない．しかし，最適化問題を解く際に導入される正則化項によって，ゆう度比を保守的に見積もる方式になる．正則化項は標本空間から要素が得られない状況で，推定されるゆう度比が標本空間の全体で等しいという事前知識を与える．これにより，ゆう度比が局所的に高くなりすぎることを防ぐ．また，連続的な標本空間の構造を捉える目的で基底関数が用いられる．そのため，基底関数の選択が重要となる．

以上を踏まえて本論文では，離散的な標本空間からのゆう度比推定に uLSIF を応用し，保守的な推定を実現する．この際に問題となるのは基底関数の選択である．連続的な標本空間を対象とする場合，空間の構造を活用するため，基底関数の定義中でガウスカーネルがよく用いられてきた．しかし，ガウスカーネルでは確率の標本空間が離散であることを考慮できない．そこで，我々は空間ごとに独立な基底関数を用いる．この基底関数を用いて uLSIF を応用すると，最ゆう推定によって求められるゆう度比を目的関数内部の正則化パラメータによって保守的に推定する方法となることが分かった．新聞記事コーパスを用いた単純な実験により，ゆう度比を保守的に推定する提案手法の振る舞いを明らかにし，その有効性を示す．更に，提案手法を自然言語処理におけるブートストラップ法(注2)に組み込むことで実用性も示す．

(注2)：統計学でのブートストラップ法（リサンプリングの方法）とは異なり，コーパスから言語知識を獲得する半教師有り学習の手法である．

2. 関連研究

最ゆう推定を用いて個々の確率推定量を求め，その比を取ることは素朴なゆう度比の推定法である．しかし，**1.** で述べたように，この方法は低頻度からゆう度比を求める際に，推定量を不当に高く見積もってしまう場合がある．本節では観測頻度が低い場合について，既存の対処法を説明する．

まず，頻度に対してしきい値を設け，それ以上の頻度に限ってゆう度比を推定する方法 [3] である．しかしながら，この方法ではしきい値を下回る低頻度事象からゆう度比を推定できない．

次に，推定量にかかるバイアスを抑制する方法である．オッズ比はゆう度比と同様，二つの確率分布に基づいて推定される．そのため，オッズ比推定のためのバイアス抑制法がゆう度比推定にも応用できる．文献 [4] では，それぞれの確率推定に MUE（中位不偏推定量）を用い，それらの推定量をもとにオッズ比を推定する．文献 [5] では，確率推定に EAP（事後平均値）を用いてオッズ比を推定する．これらの方法は，個々の確率推定を工夫するアプローチのため，最終的な推定対象をゆう度比に変更するのみで応用できると考えられる．ゆう度比推定の対象として低頻度事象のみを扱う場合，推定量のバイアス抑制は有効なアプローチの一つと考えられる．しかし，高・低頻度事象の両方を扱う場合，これらの方法では，観測頻度に基づいてゆう度比推定の信頼性を考慮できない．

最後に，保守的な推定法である．文献 [6] は，PMI（自己相互情報量）の推定において，対数内部のゆう度比計算に信頼区間を使用することを提案した．具体的には，ゆう度比の分母を構成する確率分布の推定量に信頼区間の上限，分子を構成する確率分布の推定量に信頼区間の下限を採用する．この PMI は保守的な推定量となり，提案手法と似た性質を示す．この手法は二つの信頼区間を使用するため，パラメータとして二つの信頼係数をもつ．また，文献 [7] は，ペナルティ項を使用して PMI の推定量をディスカウントする手法を提案している．この手法はパラメータをもたないが，対数外部にペナルティ項があるため，PMI の推定のみに使用できる．条件付き確率推定でも保守的な推定法が提案されている．条件付き確率は二つの確率の比で表現できるため，ゆう度比の特殊な場合と考えることもできる．文献 [8] は，確率推定量に信頼区間の

下限を使用し，保守的な推定を実現する．文献 [9] は，条件付き確率を観測頻度の比によって最ゆう推定し，分母の頻度に正の定数を加算する．この手法は本論文の提案手法と似た推定法であるが，推定対象が条件付き確率に限定されており，導出過程も異なる．提案手法は，条件付き確率を含むゆう度比全般の推定に応用可能なため，文献 [9] の手法をゆう度比推定へと一般化したものと考えることもできる．

3. ゆう度比の直接推定法

ゆう度比の直接推定法は，ロジスティック回帰による方法 [10]，カルバック・ライブラー情報量を用いた方法 [11]，最小 2 乗法による方法 [2] などがある．本節では，最小 2 乗法による方法である uLSIF（unconstrained Least-Squares Importance Fitting）を説明する．

データの定義域を $D\,(\subset \mathbb{R}^d)$ で表す．$\mathbb{R}^d$ は実 d–次元空間である．いま，確率密度 $p_{\mathrm{de}}(x)$ をもつ確率分布に独立に従う i.i.d. 標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$，及び $p_{\mathrm{nu}}(x)$ をもつ確率分布に独立に従う i.i.d. 標本 $\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ が与えられたとする．ただし，$p_{\mathrm{de}}(x)$ は

$$p_{\mathrm{de}}(x) > 0 \text{ for all } x \in D$$

を満たすと仮定する．本節では，二組の標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ と $\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ から次のゆう度比を直接推定する問題を扱う．

$$r(x) = \frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}$$

uLSIF では，推定するゆう度比 $r(x)$ を次の線形モデルで表現する．

$$\hat{r}(x) = \sum_{l=1}^{b} \beta_l \varphi_l(x) \tag{1}$$

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_b)^{\mathrm{T}}$ は標本から学習されるパラメータ，$\{\varphi_l(x)\}_{l=1}^{b}$ は非負値を取る基底関数である．b 及び $\{\varphi_l(x)\}_{l=1}^{b}$ は標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$，$\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ と独立である．この手法では次の拘束無し最適化問題を解く(注3)．

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^b} \left[\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{H}}\boldsymbol{\beta} - \hat{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{\beta} + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right] \tag{2}$$

ここで，$\hat{\boldsymbol{H}}$ は (l, l') 番目の要素 $\hat{H}_{l,l'}$ をもつサイズ $b \times b$ の行列である．要素 $\hat{H}_{l,l'}$ は次のように定義される．

$$\hat{H}_{l,l'} = \frac{1}{n_{\mathrm{de}}} \sum_{i=1}^{n_{\mathrm{de}}} \varphi_l(x_i^{\mathrm{de}})\varphi_{l'}(x_i^{\mathrm{de}}) \tag{3}$$

$\hat{\boldsymbol{h}}$ は l 番目の要素 $\hat{h}_l$ をもつ b 次元ベクトルである．要素 $\hat{h}_l$ は次のように定義される．

$$\hat{h}_l = \frac{1}{n_{\mathrm{nu}}} \sum_{j=1}^{n_{\mathrm{nu}}} \varphi_l(x_j^{\mathrm{nu}}) \tag{4}$$

式 (2) では $\boldsymbol{\beta}$ に対する正則化のため，ペナルティ項 $\frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$ を導入する．λ (≥ 0) は正則化パラメータ，$\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}/2$ は l_2–正則化項である．この式は拘束無し凸二次計画問題であり，その解は次式で解析的に計算できる．

$$\tilde{\boldsymbol{\beta}}(\lambda) = (\hat{\boldsymbol{H}} + \lambda \mathbf{1}_b)^{-1}\hat{\boldsymbol{h}}$$

$\mathbf{1}_b$ は要素が全て 1 の b 次元ベクトルである．式 (2) ではパラメータの非負制約がないため，幾つかのパラメータは負の値となることが考えられる．そこで，ゆう度比 $r(x)$ の非負性を考慮し，解を修正する．

$$\hat{\boldsymbol{\beta}}(\lambda) = \max(\mathbf{0}_b, \tilde{\boldsymbol{\beta}}(\lambda))$$

上式の ‘max’ 操作はベクトルの要素ごとに適用される．$\mathbf{0}_b$ は要素が全て 0 の b 次元ベクトルである．この $\hat{\boldsymbol{\beta}}(\lambda)$ が uLSIF の解となる．

uLSIF では，ゆう度比の直接推定に標本空間の構造を利用する．そのため，基底関数を使用し，連続的な標本空間から得た標本をもとにゆう度比を推定する．また，過学習を防ぐために l_2–正則化項を導入する．この正則化項は標本空間から要素が得られない状況下で，パラメータが一様という事前知識を与える．

4. 提 案 手 法

uLSIF の原論文 [2] では，基底関数の定義中でガウスカーネルを用いた．しかし，本論文で標本の要素に相当するものは連続的な標本空間から得られる実数値ではなく，離散的な標本空間から得られる単語，バイグラムなどである．ガウスカーネルでは標本空間が離散であることを考慮できない．離散的な標本空間を扱うために，文字列カーネルや木カーネルの使用が考えられるが，これらのカーネルを使用すると推定モデル

(注3)：式 (2) の導出過程は uLSIF の原論文 [2] を参照のこと．

の定式化やゆう度比の効率的な推定が難しい．そこで，次の単純な基底関数 $\{\varphi_l(x)\}_{l=1}^{v}$ を用いる．ここで，x は単語，バイグラムなどの要素，v は存在しうる要素の種類数であり，要素の種類ごとに対応する基底関数を一つ定義する．この基底関数はクロネッカーのデルタと考えることもできる．

$$\varphi_l(x) = \begin{cases} 1 & (x = w_{(l)}) \\ 0 & (x \neq w_{(l)}) \end{cases} \tag{5}$$

この基底関数は異なる要素間の関係を捉えられないが，uLSIF と組み合わせると解析的に解が求められる．加えて，導出される推定式が単純で扱いやすい利点がある．添え字 l は v 種類存在する要素から，特定の要素を指定する．すなわち，$w_{(l)}$ は v 種類ある要素のうち，l 種類目の要素を指す．式 (5) の基底関数を uLSIF の枠組みに当てはめる．推定対象とするゆう度比 $r(x)$ を次の線形モデルで表現する．

$$\begin{aligned} \hat{r}(x) &= \sum_{l=1}^{v} \beta_l(\lambda)\varphi_l(x) \\ &= \beta_l(\lambda) \end{aligned}$$

ただし，$x = w_{(l)}$ とする．このモデルは式 (1) で定義した線形モデルに対応する．

式 (5) より，基底関数 $\varphi_l(x)$ は，x が $w_{(l)}$ と等しくないときに 0 となる．よって，式 (3) 及び式 (4) に対応する $\hat{H}_{l,l'}$，$\hat{h}_l$ は次式となる．

$$\begin{aligned} \hat{H}_{l,l'} &= \begin{cases} \frac{1}{n_{\mathrm{de}}} \sum_{i=1}^{n_{\mathrm{de}}} \varphi_l(x_i^{\mathrm{de}})\varphi_{l'}(x_i^{\mathrm{de}}) & (l = l') \\ 0 & (l \neq l') \end{cases} \\ &= \frac{1}{n_{\mathrm{de}}} c_{\mathrm{de}}(w_{(l)}) \quad (l = l') \\ \hat{h}_l &= \frac{1}{n_{\mathrm{nu}}} \sum_{j=1}^{n_{\mathrm{nu}}} \varphi_l(x_j^{\mathrm{nu}}) \\ &= \frac{1}{n_{\mathrm{nu}}} c_{\mathrm{nu}}(w_{(l)}) \end{aligned}$$

要素 $\hat{H}_{l,l'}$ をもつ $v \times v$ 行列 $\hat{\boldsymbol{H}}$ は対角成分のみが残り，それ以外の要素は 0 となる．$c_*(w_{(l)})$ は，確率密度 $p_*(w_{(l)})$ をもつ確率分布から観測した $w_{(l)}$ の頻度である．

前述したように，行列 $\hat{\boldsymbol{H}}$ には対角成分のみが残る．よって，$x = w_{(l)}$ のとき，推定量 $\hat{r}(x)$ は次式となる．

$$\begin{aligned} \hat{r}(x) &= \tilde{\beta}_l(\lambda) \\ &= (\hat{H}_{l,l} + \lambda)^{-1}\hat{h}_l \\ &= \left(\frac{1}{n_{\mathrm{de}}} c_{\mathrm{de}}(w_{(l)}) + \lambda\right)^{-1} \cdot \frac{1}{n_{\mathrm{nu}}} c_{\mathrm{nu}}(w_{(l)}) \end{aligned} \tag{6}$$

uLSIF では，負の値となるパラメータを 0 に丸める必要がある．しかし，式 (6) は非負であり，パラメータの推定量 $\tilde{\beta}_l(\lambda)$ がそのまま $r(x)$ の推定量となる．式 (6) は単純な形をしており，正則化パラメータ λ が推定量を保守的に見積もる効果を生む．この推定量は λ が 0 のとき，ゆう度比の分母・分子にあたる確率分布の最ゆう推定量(注1)をそれぞれ求め，それらの比を取った結果に等しい．提案手法はゆう度比の分母・分子をそれぞれ補正するのではなく，分母のみを補正するという特徴をもつ．

5. 比較手法

本論文では二つの実験を行う．**6.** では，新聞記事コーパスからカタカナ語直前に出現することのあるバイグラムを予測する実験により，提案手法の振る舞いを調査し，その有効性を示す．**7.** では，ブートストラップ法を用いて日本語の科学ニュース記事から雑誌名を抽出する実験により，提案手法の実用性を示す．両実験とも低頻度事象を多く扱うため，低頻度からゆう度比をどのように推定するかで実験で得られる性能が大きく変化する．両実験の共通目的は，観測頻度の低さに応じてゆう度比を保守的に推定する，提案手法の効果を検証することである．そこで，低頻度事象の扱い方が異なる既存手法を選び，提案手法の比較対象とした．

二種類の試行に対し，表 2 に示す集計表を考えたとき，推定対象のゆう度比を次式で定義する．

$$\mathrm{LR}(x) = \frac{p_1}{p_2} \tag{7}$$

n_t，k_t，p_t は，試行 t ($\in \{1, 2\}$) の試行回数，成功回数，成功確率を示す．なお，p_t の最ゆう推定量 $\hat{p}_t$ は k_t/n_t で表される(注1)．手法 1 は，ベースラインとな

表 2　二値出力をもつ試行に対する集計表
Table 2　The tabulation table of binary outcomes trials.

	Outcome		
Trial	Success	Failure	Total
1	k_1	$n_1 - k_1$	n_1
2	k_2	$n_2 - k_2$	n_2

る単純な推定法である．手法 2 と手法 3 は，推定量にかかるバイアスを抑制する．手法 4 は，観測頻度の低さに応じてゆう度比を保守的に推定する．

手法 1：最ゆう推定量を用いた方法（MLE）

p_t を最ゆう推定量(注1)によって推定し，その比を取る．この方法は単純だが，低頻度からゆう度比を求めるとき，その推定量を不当に高く見積もる場合がある．

手法 2：事後平均値を用いた方法（EAP）

p_t に関する事前分布としてベータ分布 $\beta(a_t, b_t)$ を仮定する．そして，事後分布の平均値（EAP: Expected A *Posteriori*）を p_t の推定量として求め，その比を取る．本手法は，低頻度の観測事象からゆう度比を推定する場合に有効とされている [5]．本手法の原論文では，ハイパーパラメータ a_t と b_t を最ゆう推定によって求めている．しかし，本論文では原論文と異なり，n_t が大きい．この場合，原論文の方法では最ゆう推定解を解析的に求めることが困難である．それゆえ，データの平均と分散を求め，ベータ分布のそれに一致するようにハイパーパラメータを推定する．

手法 3：Modified MUE を用いた方法（MMUE）

p_t を中位不偏推定量（MUE: Modian Unbiased Estimator）によって推定し，その比を取る．本手法も手法 2 と同様，低頻度の観測事象からゆう度比を推定する場合に有効とされている [4]．

手法 4：信頼区間を用いた方法（CI）

p_1 と p_2 の推定に信頼区間の下限，上限をそれぞれ用いる．本手法は，ゆう度比を保守的に推定する [6]．信頼区間の構築には，二項分布を正規近似する方法がよく用いられる．しかし，正規近似が有効に機能する状況は二項分布の母平均が 0.5 に近い場合であり，本論文で扱うのは母平均が 0 に近い場合である．そこで，信頼区間の両端が 0 や 1 に近い場合に有効とされる Wilson score interval [12] を用いる．本手法は二つの信頼区間を構築する必要があり，二つの信頼係数をパラメータとしてもつ．信頼係数を別々に変化させて最適値を探すこともできるが，提案手法と対等な比較のため，二つの信頼係数に同じ値を採用する．

6. 提案手法の有効性検証

本節では，提案手法の振る舞いを調査し，ゆう度比を保守的に推定することの有効性を検証する．実験ではゆう度比を用いて，カタカナ語直前に出現することのある文字バイグラムを予測する．文字列「東南アジア」を例にすると，カタカナ語が「アジア」，その直前にあるバイグラムが「東南」となる．この実験を選んだ理由は次の三点である．第一に，この実験は単純な問題設定のため，提案手法の振る舞いを調査しやすい．第二に，カタカナ語直前に現れるバイグラムは定まっており，一意な正解が定義できる．つまり，バイグラムの予測性能を定量的に測定できる．第三に，バイグラムの頻度分布はべき乗則に従い，低頻度のバイグラムを多く扱うためである．この理由から，低頻度の扱い方によって予測性能が大きく変化すると考えられ，低頻度の扱い方が異なる手法間の性能差を観察しやすい．加えて，低頻度から推定されるゆう度比を高く見積もると，予測性能の大幅な低下が想定される．そこで，保守的な推定法を用いて性能向上を確認する．

表 3　データ集合に含まれるバイグラム
Table 3　Bigrams contained in each dataset.

年版	データ集合	全体		カタカナ語の直前	
		種類数	総頻度	種類数	総頻度
91	Training	223,538	4,728,204	13,805	88,594
	Validation	71,474	435,371	3,227	8,403
	Test	72,248	455,855	3,372	8,795

6.1 実験手順

実験では，1991 年版の毎日新聞コーパス(注4)から作成したデータ集合を使用する．データ集合は次の手順で作成する．コーパスからランダムに 12,000 件の記事を抽出し，それらを訓練データ（10,000 記事），バリデーションデータ（1,000 記事），テストデータ（1,000 記事）と割り振る．データ集合に含まれるバイグラムの種類数・総頻度を表 3 に示す．

実験は次の流れで行う．まず，訓練データから文字バイグラムの頻度を学習する．訓練データに含まれる全バイグラムについて，カタカナ語直前，訓練データ全体での出現頻度を種類別に数え上げる．これらの頻度は，表 2 で示す k_1，k_2 に対応する．また，カタカナ語直前，訓練データ全体に現れる全バイグラムの総頻度は n_1，n_2 にそれぞれ対応する．次にテストデータに含まれる任意バイグラム w について，提案手法及び **5.** の手法を用いて次のゆう度比を推定する．

$$\mathrm{LR}(w) = \frac{p(w \mid O_k)}{p(w)} \tag{8}$$

O_k はバイグラムがカタカナ語直前に出現することを

(注4)：実験のランダム性を担保するため，91 年版の他にも，92 年版から 94 年版のコーパスを用いて年版ごとに同様の実験も行った．同様の実験結果が得られたため，紙面の都合上，91 年版の実験のみを掲載する．

意味する．$p(w \mid O_k)$，$p(w)$ はカタカナ語直前，訓練データ中の任意位置での w の出現確率であり，式 (7) の p_1，p_2 にそれぞれ対応する．ゆう度比の推定には学習した頻度を使用する．手法ごとに，推定したゆう度比の降順に文字バイグラムを並べてランク付けし，上位から正誤判定をする．判定対象のバイグラムがテストデータ中でカタカナ語直前に 1 回でも出現すれば正解，それ以外は不正解とする．最後に，ランク上位 8,000 件についてランク—再現率曲線を描く．この曲線は横軸をランク，縦軸をそのランクまでの再現率とした曲線であり，原点と曲線の一点を結んだ直線の傾きが適合率に比例する．再現率と適合率は次式で定義される．

$$\text{Recall} = \frac{|\{w \mid w \in R\}|}{|R|},$$

$$\text{Precision} = \frac{|\{w \mid w \in R\}|}{|\{w\}|}$$

ここで，w は正解判定の対象バイグラム，R はテストデータ中でカタカナ語直前に出現したバイグラムの集合であり，正解集合を意味する．

EAP と MMUE は，訓練データから学習した頻度を用いて確率ごとにハイパーパラメータを推定した．提案手法と CI は次のようにパラメータを推定した．パラメータを変化させたそれぞれの場合について，バリデーションデータをテストデータとみなし，上記と同手順でランク—再現率曲線を描いた．そして，曲線下面積が最大となるパラメータを最適値として採用した．提案手法は正則化パラメータ λ を 10^{-9} から 10^{-1} まで 10 倍ずつ増加させ，10^{-2} を最適値として採用した．CI は信頼区間の幅が片側 90%から 99%まで，1%ずつ増加するように信頼係数を変化させた．結果として片側 99%の場合が最適値となったため，信頼区間の幅が片側 99.0%から 99.9%となるよう，0.1%ずつ信頼係数を変化させ，片側 99.9%を最適値として採用した．

6.2 実験結果

ランク—再現率曲線を図 1 に示す．この曲線は横軸をランク，縦軸をそのランクまでの再現率とし，原点と曲線の一点を結んだ直線の傾きが適合率に比例する．同一のランクにおいて，縦軸の最も大きい（つまり，再現率の最も高い）手法がそのランクにおいて最も優れた性能をもつ．

図 1 から分かるように，MLE はランク再現率曲線が直線に近い形状である．これは，ランクに対してほぼ一定の割合で正解を発見したことを意味する．しかし本来は，ランク上位ではカタカナ語の直前に現れやすいバイグラムが位置することが望ましく，上位での適合率向上を阻害している要因があると考えられる．EAP と MMUE はランクのごく上位で適合率が向上するが，それ以降はほぼ横ばいの直線である．一方，提案手法は全体的に最良の性能であり，CI は提案手法にわずかに劣るがほぼ同等の性能を有している．これらの手法はランク上位での適合率が高く，ランク下位でも徐々に再現率が向上している．これはランク上位で多くの正解を発見し，ランク下位でも継続的に正解を発見できたことを意味する．

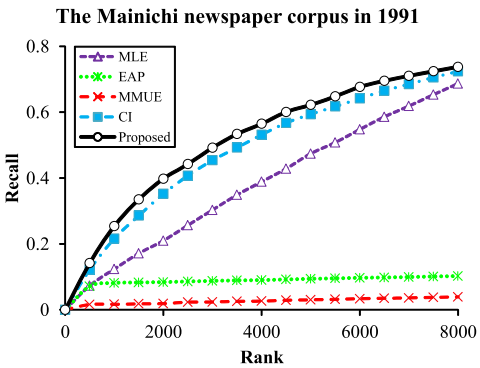


図 1　ランク—再現率曲線
Fig. 1　Rank-recell curves.

以上では，ランク—再現率曲線から各手法の性能を比較した．次に，それぞれの性能がもたらされた要因を明らかにするため，各手法の振る舞いを定性的に分析する．各手法がランク付けしたバイグラムの一例を表 4 に示す．ランク外はバイグラムが 8,000 位より下位であることを意味する．

MLE，EAP，MMUE は，訓練データ全体の頻度に占めるカタカナ語直前の出現割合を重視する傾向にある．この働きにより,「東南」などのバイグラムに加え,「粗大」といった低頻度の正解も得られる．しかし，高い頻度をもつ「東京」などのバイグラムが下位となり,「認証」などの不正解が上位の多数を占める．結果として，上位での適合率が低くなったと考える．また，EAP と MMUE は「冒涜」など，訓練データに出現しないバイグラムのゆう度比を不当に高く見積もってしまう(注5)．この問題が全体的な再現率の低下を招いたと考える．

(注5)：MLE では推定値の分母がゼロになってしまい，ゆう度比を計算できない．そのため，この場合の推定値をゼロとして扱った．

表 4　ランク付けしたバイグラムの例
Table 4　Examples of ranked bigrams.

バイグラム	出現頻度（訓練データ）		各手法におけるランク					正誤
	全体	カナ直前	MLE	EAP	MMUE	CI	Proposed	
東南	127	121	90	90	ランク外	4	83	○
東京	3,941	104	6,623	ランク外	ランク外	1,657	113	○
粗大	2	2	1	62	19	1,370	4,226	○
認証	1	1	1	1	28	3,391	5,912	×
冒涜	0	0	ランク外	585	89	ランク外	ランク外	×

提案手法は,「東南」や「東京」といった全体で高頻度かつカタカナ語直前にも現れやすいバイグラムがランク上位となる傾向がある．よって，このようなバイグラムを上位に位置づける働きが，ランク上位での高い適合率をもたらしたと考える．ランク下位では,「粗大」,「認証」などの訓練データ全体であまり出現しないが，カタカナ語直前で出現することのあるバイグラムが位置付けられる．このようなバイグラムは，不正解が多いものの，テストデータ中でもカタカナ語直前に来る可能性がある．このようなバイグラムをランク下位とすることで，下位でも正解を発見でき，高い再現率を維持できたと考える．以上から，提案手法は高頻度と低頻度のバイグラム両方を有効に扱う方法であることが示唆された．

CI は提案手法と似た性質をもっている．ただし,「東京」と「粗大」の順位関係が提案手法と逆転していることに注意する．これは，提案手法が訓練データ全体での出現頻度を重視するのに対し，CI は全体的な出現頻度に占めるカタカナ語直前での出現割合の高いバイグラムを重視する傾向にあることを示唆している．

7.　アプリケーションでの実用性検証

6. で実施した実験に加えて，提案手法の実用性を示すために，自然言語処理におけるブートストラップ法に提案手法を組み込む．ブートストラップ法は，コーパスから語彙知識を獲得する半教師有り学習の手法であり，ウェブページ分類 [13]，語義曖昧性解消 [14], [15]，固有表現抽出・分類 [16], [17]，構文解析 [18] や情報抽出 [19] など様々なタスクに応用される．この方法は，シードと呼ばれる少数の教師データをもとにコーパスから文脈パターンを抽出し，それを手がかりとしてインスタンス（シードと同じドメインに属する獲得対象）を抽出する．そして，抽出したインスタンスをシードに加えて同様の処理を繰り返す．ブートストラップ法では，過去に抽出したパターンをインスタンスの抽出に再利用するため，誤ったパターンは誤ったインスタンスの増加を招く．それゆえ，繰り返しごとに確実なパターン・インスタンスの抽出が求められる．そこで，提案手法を用いてパターンをスコアリングし，充分な頻度でインスタンスと共起するパターンを優先してインスタンス抽出に使用する．

実験では，日本語の科学ニュース記事から科学雑誌名を抽出する．インスタンスのドメインを科学雑誌名とした理由は次の二点である．第一に，雑誌名は種類が豊富で，様々な表記法があるためである．和名表記の雑誌名は，記事の執筆者によって多様な表記ゆれが生じる．例えば，雑誌名「Cell Metabolism」は「セル・メタボリズム」,「セルメタボリズム」,「細胞代謝」といったように表記ゆれが生じる．この場合，多くの雑誌名を包含する大規模な教師データの用意は困難であり，ブートストラップ法などの半教師有り学習法の使用が適している．第二に，記事内の雑誌名は役割がおおむね定まっており，特定の文脈パターンで出現しやすいためである(注6)．この性質ゆえ，ゆう度比で文脈パターンを適切に捉えられたか否かによってインスタンスの抽出性能が変化し，手法間の性能差が観察しやすい．筆者らの調査では，和名を含む科学雑誌名の抽出は研究されておらず，雑誌名を列挙したリストも公開されていないことを確認した．そのため，抽出した雑誌名のリストを公開する(注7)．

7.1　ブートストラップ法のアルゴリズム

以下の手順を繰り返し，インスタンスを抽出する．

(1) パターン抽出：シードと共起する文脈パターンを全て抽出し，種類ごとに出現頻度を記録する．また，シードと共起する全パターンの総頻度も記録する．任意文字列（2 文字から 50 文字）と共起するパターンを全て抽出し，シードとも共起するパターンについて種類ごとに頻度を記録する．同時に，任意文字列と共起する全パターンの総頻度も記録する．各頻度は，**5.** の表 2 で示す k_1，n_1，k_2，n_2 に対応する．雑誌名は

(注6)：この性質は予備実験で確認した．
(注7)：https://doi.org/10.5281/zenodo.2562355

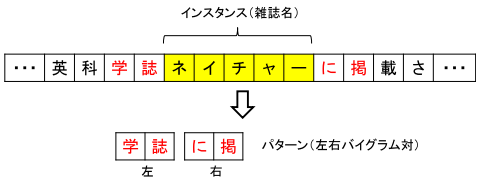


図 2　パターンの抽出
Fig. 2　Extraction of the pattern.

括弧で囲まれることが多く，左右文脈での対応が重要となる．そのため，左右の文字 N グラムを個別に用いるのではなく，それらを組み合わせた N グラム対をパターンとする．ただし，N グラム対は個々の N グラムよりも頻度が疎になるため，扱いに注意を要する．インスタンスと共起するパターンの抽出例を図 2 に示す．

(2) パターンのスコアリング：任意パターン θ ごとに手順 (1) で求めた頻度を用いて，次のゆう度比を推定する．

$$\mathrm{Score}(\theta) = \frac{p(\theta \mid O_j)}{p(\theta)} \tag{9}$$

O_j はパターンが雑誌名と共起することを示す．$p(\theta \mid O_j)$ は θ が雑誌名と共起する確率，$p(\theta)$ は θ が任意文字列と共起する確率である．それぞれの確率は，式 (7) の p_1，p_2 に対応する．

(3) インスタンス抽出：スコア降順からパターンを用いてインスタンスを抽出する．インスタンスの長さは雑誌名か否かの判断が難しい 1 文字を除いた 2 文字から 50 文字とする．なお，シードに含まれるもの及び過去に抽出したものは抽出しない．

(4) インスタンス選択：抽出したインスタンスのうち，パターン抽出に有効なもののみを残してシードに加える．インスタンスの選択法は **7. 2** の実験条件で述べる．

7. 2　実 験 条 件

科学雑誌名を含む可能性の高い日本語ニュース記事をコーパスとして使用する．具体的には，複数のニュースサイトから過去およそ 10 年分のニュース記事を収集し，「学誌 OR 論文誌 OR 学術誌」という検索条件で絞り込んだ 30,076 記事を使用する．インスタンス抽出の際は記事本文のみを参照する．シードインスタンスを表 5 に示す．抽出元が日本語記事であるため，雑誌の和名と英名が多く抽出される．大別した雑誌名の表記を表 6 に示す．実験では，**7. 1** の各手順を 5 回繰り返し，手順 (3) ではスコア降順 1,000 件のインスタンスを抽出する．また，同じインスタンスであっても共起するパターンによってスコアの推定値が異なる．同名のインスタンスで複数のスコアがある場合は，最高のスコアをそのインスタンスのスコアとみなす．式 (9) のゆう度比を推定する手法として提案手法及び **5.** で述べた手法を用いる．

表 5　シ　ー　ド
Table 5　Seeds.

Scientific Reports
サイエンティフィック・リポーツ
サイエンティフィック・リポーツ（Scientific Reports）
サイエンティフィックリポーツ
サイエンティフィックリポーツ（Scientific Reports）
PLOS ONE
プロス・ワン
プロス・ワン（PLOS ONE）
プロスワン
プロスワン（PLOS ONE）

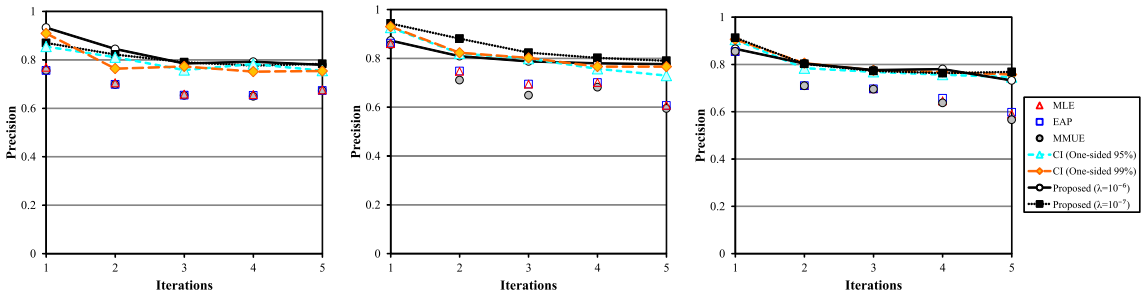
表 6　雑誌名の表記
Table 6　Notations of journal names.

英名
Neuron, Cell Research
和名
ニューロン，セル・リサーチ
英名・和名の併記
ニューロン（Neuron），セル・リサーチ（Cell Research）
補足情報付き
ニューロン電子版，セル・リサーチ（電子版）

以下の $3 \times 2 = 6$ 条件について，ブートストラップ法を適用し，雑誌名の抽出性能を測定する．

パターンの長さ（3 種類）：インスタンスと共起するバイグラム対，トライグラム対，4 グラム対．

インスタンスの選択法（2 種類）：人手でラベル付けした雑誌名のみ，スコアの降順 1,000 件のインスタンス．適切なパターンの長さを決めることは困難である．パターンを長く取れば，インスタンスの出現文脈をより正確に捉えられるが，低頻度のパターンが増加する．そこでパターンの長さをバイグラム対，トライグラム対，4 グラム対と変化させた場合について，抽出性能の変化をパターンの頻度に着目して考察する．シードに加えるインスタンスの選択法については，二つの方法を試す．ブートストラップ法の性能低下の原因として，誤ったパターン・インスタンス抽出による影響がある．実験では，パターンのスコアリングを工夫し，抽出性能の低下抑制を試みるが，誤ったインスタンスによるパターンの誤抽出は防止できない．そこで，人手でラベル付した正しい雑誌名のみをパターン抽出に利用し，ゆう度比の推定方法のみに依存するブートス



(a) Bigram pair　(b) Trigram pair　(c) Four-gram pair

図 3　繰り返しごとの適合率．人手でラベル付した雑誌名のみをパターンの学習に用いた．
Fig. 3　Precisions for each iteration. For pattern learning, we only used journal names labeled by hand.

トラップ法の性能変化を観測する．ただし実環境では手続き自動化の観点から，スコア上位となるインスタンスをそのまま用いることが多い．そのため，スコアの降順 1,000 件のインスタンスをシードに加える実験も行い，提案手法のロバスト性を評価する．

ブートストラップ法の抽出性能は繰り返しごとに累積の適合率で測定する．インスタンスの正誤判定は人手で実施し，正解したインスタンスには雑誌名のラベルを付与する．適合率の定義を示す．

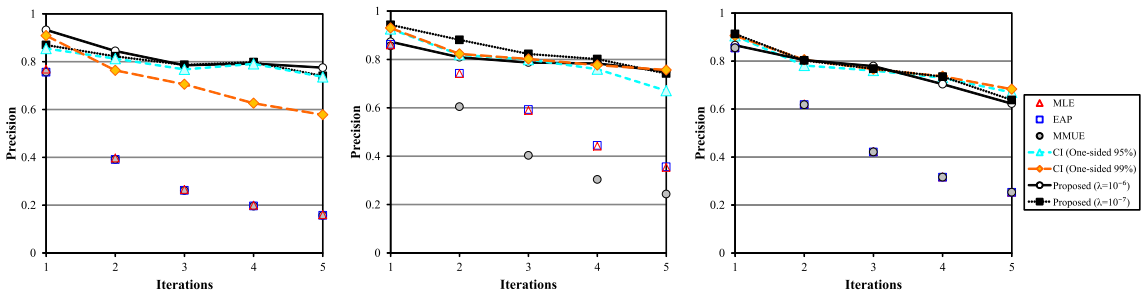
$$\text{Precision} = \frac{|\text{正解したインスタンス}|}{|\text{抽出したインスタンス}|}$$

各手法のパラメータ設定方法を述べる．EAP と MMUE は，**7.1** の手順 (1) で求めた頻度を用いて確率ごとにハイパーパラメータを推定した．CI は二つの信頼区間がともに片側 95%，99%となるように信頼係数を設定した．これらの区間幅は一般的によく用いられる．提案手法は，λ の値を 10^{-9} から 10^{-1} まで 10 倍ずつ変化させ，初期のシードから求めた頻度をもとに各パターンのゆう度比を推定した．λ を 10^{-8} とすると，コーパス全体で出現頻度が 10 未満のパターンが推定値の降順上位 10 件に含まれた．この場合，λ の値が小さすぎるため，低頻度から推定されるゆう度比を不当に高く見積もったと考える．一方，λ を 10^{-5} とすると低頻度の影響を低減できるが，コーパス全体での頻度が大きく異なるにもかかわらず，雑誌名との共起頻度が近いパターンが類似した推定値となった．これは，λ の値が大きすぎるゆえ，頻度を過剰評価したと考える．以上より，λ の最適値は 10^{-6} あるいは 10^{-7} 付近と予想し，この二つをパラメータの値とした．なお，EAP と MMUE はブートストラップ法の繰り返しごとにパラメータを推定するが，CI と提案手法は繰り返しの初回でパラメータを決め，以降はそれを固定して用いた．

7.3　実験結果

パターン抽出に人手でラベル付した雑誌名のみを使用した結果を図 3，スコアの降順 1,000 件のインスタンスをシードに加えた結果を図 4 に示す．各グラフについて，横軸は繰り返しの回数，縦軸はその回数までの累積適合率を表す．高い適合率をもつ手法が性能の良い手法である．

図 3 より，MLE，EAP，MMUE はパターンを長く取ると，バイグラム対と比較して途中まで良い適合率を維持している．しかし，トライグラム対の場合は 5 回目で適合率が大きく減少した．適合率減少の原因は，多義性のある曖昧な雑誌名（例えば「RNA」）から，雑誌名と無関係な低頻度パターンを学習し，誤ったインスタンスを大量抽出したためである．この現象は意味ドリフトと呼ばれる．パターンを 4 グラム対にすると，意味ドリフトの影響を 4 回目から受け，適合率減少のタイミングが早まった．よって，これらの手法は正しい雑誌名のみからパターンを抽出しても低頻度の悪影響を強く受け，全体的に低い適合率を示したと考える．保守的な推定手法（提案手法と CI）は MLE，EAP，MMUE よりも全体的に高い適合率を維持している．また，提案手法とはパターンをバイグラム対ではなくトライグラム対とすると最良の性能を示した．これは，提案手法が低頻度パターンの悪影響を避けつつ，雑誌名の出現文脈を正確に捉えたことを意味する．以上から，提案手法と CI は低頻度のパターンを保守的に見積もり，多くの雑誌名を抽出できたと考える．



(a) Bigram pair　(b) Trigram pair　(c) Four-gram pair

図 4　繰り返しごとの適合率．高いスコアをもつ上位 1,000 件をパターンの学習に用いた．
Fig. 4　Precisions for each iteration. For pattern learning, we used top 1,000 instances which have high scores.

図 4 と図 3 について，パターンの長さが同じ結果どうしを比較する．MLE，EAP，MMUE は不正解の（雑誌名ではない）インスタンスをパターン抽出に利用すると，繰り返しごとの適合率が大きく低下する．不正解のインスタンスからは低頻度の無関係なパターンが大量に抽出される．よって，これらの手法は増加した低頻度パターンに対処できず，性能が大きく低下したと考えられる．ブートストラップ法では，スコア上位となるインスタンスをパターン抽出にそのまま用いることが多い．そのため，これらの手法は実際のブートストラップ法では有効に機能しないことが示唆された．提案手法と CI は繰り返しの 5 回目まで，適合率に大きな変化は見られない．つまり，これらの手法は低頻度のパターンが大量に混入しても多数の雑誌名を抽出できている．この結果から，提案手法と CI は低頻度パターンの混入に対してロバストであり，実環境でも有効に機能すると考えられる．

8. 考　　察

データを確率的に取り扱う問題において，統計的尺度の推定は手法の構成やデータ分析の基盤的役割を担う．本論文では，統計的尺度の一つであるゆう度比について，保守的な推定法を提案した．実験では，提案手法が比較手法 CI と似た性質をもち，ブートストラップ法による雑誌名抽出タスクにおいて同等の抽出性能を獲得しうることを示した．CI は二つの信頼係数をパラメータとするヒューリスティックである．この方法は扱うデータに応じて，信頼区間の推定法を適切に仮定する必要がある．提案手法は，ゆう度比の直接推定法を確率の標本空間が離散の場合に応用したものである．我々の手法はゆう度比の推定誤差を最小にする意味付けがあり，調節すべきパラメータが一つでよい利点がある．加えて，推定式も単純で扱いやすい．

6. と **7.** の両実験で提案手法が有効であった理由の一つとして，ゆう度比を構成する分母の確率のみで，ゆう度比推定の不確かさを考慮できたことが考えられる．例えば，**6.** の式 (8) を推定する際，バイグラム w がカタカナ語直前に現れる頻度 k_1，訓練データ全体に現れる頻度 k_2 の間に $k_1 \leq k_2$ が成立する．すなわち，k_2 が低頻度であれば，k_1 は更に低頻度となる．それゆえ，分母の確率推定量のみに着目し，正則化パラメータで補正することで保守的な推定が実現できる．

k_1 と k_2 が無関係な場合には保守的な推定が機能しない状況もある．この状況では，推定の不確かさは分母・分子の片方のみでは考慮できないと考えられる．このときの合理的な推定法は検討の余地があるものの，提案手法が保守的に推定できるゆう度比はよく用いられている．例えば，PMI（自己相互情報量）は二つの事象間の関連度合いを測る尺度である．PMI を用いるにはゆう度比を推定する必要があり，この推定に提案手法を使用できる．また，リフトと呼ばれるゆう度比はパターンマイニングで使用され，リフトの推定にも提案手法が有効と考えている．これらの尺度は様々な研究に用いられており，提案手法の応用先は広いと考えている．低頻度事象の扱いは古くから議論される問題であり，高・低頻度の事象は別々に扱われることが多い．提案手法は両方の頻度を扱うため，高・低頻度を同時に扱う研究で有効と考えている．

確率推定量を補正するために，確率のスムージング法が多く提案されている [20]～[28]．スムージング法

は確率推定のために用いられ，ゆう度比を扱う提案手法とは異なるものである．しかし，スムージング法をゆう度比推定に応用することで，形式的に提案手法と類似の推定式が導出される可能性もある．そこで，ゆう度比を構成するそれぞれの確率分布から推定量を求める際に，スムージング法の活用を検討する．ここでは，ゆう度比推定にスムージング法を用いた場合と提案手法を用いた場合の差異，及び提案手法を用いる利点について考察する．

スムージング法は，ゆう度比の分母・分子にそれぞれ適用できる．このとき，ゆう度比を構成する分母・分子の確率分布から得られる推定量がそれぞれ補正される．一方で，提案手法はゆう度比の分母にのみ，正則化項由来の定数項が出現する．また，スムージング法は観測事象から計算される確率推定量を割引いて，それを未観測事象から計算される確率推定量に分配する．これにより，未観測事象の確率推定値がゼロとなることを防ぐ．この作用のために，ゆう度比推定にスムージング法を用いると，未観測事象から推定されるゆう度比は正の値となる．それに対して，提案手法では未観測事象に基づくゆう度比の推定値をゼロとする．提案手法は全ての場合について，ゆう度比を保守的に推定することが，ゆう度比の真値との誤差を減少させることを示唆している．

前述のとおり，提案手法はゆう度比の分母のみに定数項が出現する．ゆう度比の分母となる確率分布だけに，線形補間によるスムージング法を用いれば，提案手法と類似した推定式が導出できる．線形補間を利用したスムージング法として，Jelinek-Mercer スムージング [21]，Witten-Bell スムージング [23]，Absolute ディスカウンティング [24], [25]，Dirichlet スムージング [26]，Kneser-Ney スムージング [28] などがある．これらのスムージング法は，N グラム確率を補正する目的で用いられる．また，これらのスムージング法はパラメータとして補間係数をもつが，この係数は通常，標本空間における確率推定値の総和が 1 となるように決定される．一方，提案手法では，標本空間においてゆう度比の総和を 1 とする制約はない．更に，補間係数には低次の N グラム確率が乗算され，推定対象の N グラムに応じて加算される値が異なる．提案手法では，ゆう度比の分母に加算される値は定数である．以上から，確率のスムージング法をゆう度比推定にそのまま流用しても，提案手法と同じ形式の推定式は導けない．

ここで述べたスムージング法は，一般に確率推定のために導入される．それゆえ，個々の確率分布から求められる推定量を補正しても，ゆう度比を保守的に推定できるとは限らない．特に，スムージング法を用いた場合に，未観測事象から推定されるゆう度比はゼロより大きくなることが想定される．未観測事象のゆう度比を高く見積もると，ゆう度比の推定性能を悪化させる要因となる．これは **6.** の実験結果からも明らかである．提案手法では未観測事象のゆう度比が，取り得る値の下限であるゼロとなるため，この問題を回避できる．線形補間を利用するスムージング法は，N グラム確率を補正する目的で利用され，低次の N グラムを用いる必要がある．一方，提案手法はゆう度比一般の推定に利用できる．例えば，**6.** と **7.** の実験で示したように，低次の N グラムを使用しない場合や標本空間から得られる要素が N グラム以外（**7.** の実験では N グラム対）の場合でも幅広く適用できると考えている．

9. むすび

本論文では，観測頻度に基づいて離散的な標本空間からゆう度比を推定する問題を扱った．素朴な推定方法は，ゆう度比の定義に従い，ゆう度比を構成する二つの確率分布を最ゆう推定して，その比を取ることである．しかし，低頻度からゆう度比を求めるとき，この方法は推定量を不当に高く見積もってしまう場合がある．そこで，頻度の低さに応じてゆう度比を保守的に推定する方法を提案した．提案手法は，ゆう度比の直接推定法 uLSIF の枠組みを利用し，正則化により推定量を調節する．実験では提案手法の振る舞いを明らかにし，その有効性を示した．更に，ブートストラップ法を利用した実験では，実用性も確認した．提案手法は，高頻度から推定されるゆう度比を優先的に扱い，かつ低頻度から得られる不確実性の高い推定結果も活用する研究で有効と考えている．

文　献

[1] 中西健太郎，田中利幸，上田修功，“尤度比に基づく順位づけ関数による受信者操作特性曲線下面積の漸近的性質,” 信学技報，IBISML2014-92, 2015.

[2] T. Kanamori, S. Hido, and M. Sugiyama, “A least-squares approach to direct importance estimation,” J. Machine Learning Research, vol.10, pp.1391–1445, July 2009.

[3] A. Montella, “Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types,” Accident Analysis & Prevention, vol.43, no.4,

pp.1451–1463, 2011.
[4] M. Parzen, S. Lipsitz, J. Ibrahim, and N. Klar, “An estimate of the odds ratio that always exists,” J. Computational and Graphical Statistics, vol.11, no.2, pp.420–436, 2002.
[5] K. Raweesawat, Y. Areepong, K. Jampachaisri, and S. Sukparungsee, “Odds ratios estimation of rare event in binomial distribution,” J. Probability and Statistics, pp.1–8, 2016.
[6] M. Johnson, “Confidence intervals on likelihood estimates for estimating association strengths,” Unpublished technical report, 1999.
[7] P. Pantel and D. Ravichandran, “Automatically labeling semantic classes,” NAACL-HLT, pp.321–328, 2004.
[8] 菊地真人，山本英子，吉田光男，岡部正幸，梅村恭司，“条件付き確率の保守的な推定，” 信学論（D），vol.J100-D, no.4, pp.544–555, April 2017.
[9] C. Rudin, B. Letham, and D. Madigan, “Learning theory analysis for association rules and sequential event prediction,” J. Machine Learning Research, vol.14, no.1, pp.3441–3492, 2013.
[10] S. Bickel, M. Brückner, and T. Scheffer, “Discriminative learning for differing training and test distributions,” ICML, pp.81–88, 2007.
[11] M. Sugiyama, S. Nakajima, H. Kashima, P. vonBünau, and M. Kawanabe, “Direct importance estimation with model selection and its application to covariate shift adaptation,” Advances in Neural Information Processing Systems, pp.1433–1440, 2008.
[12] E.B. Wilson, “Probable inference, the law of succession, and statistical inference,” J. American Statistical Association, vol.22, no.158, pp.209–212, 1927.
[13] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” COLT, pp.92–100, 1998.
[14] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” ACL, pp.189–196, 1995.
[15] R.F. Mihalcea and D.I. Moldovan, “A highly accurate bootstrapping algorithm for word sense disambiguation,” Int. J. Artificial Intelligence Tools, vol.10, no.1-2, pp.5–21, 2001.
[16] M. Collins and Y. Singer, “Unsupervised models for named entity classification,” EMNLP, pp.100–110, 1999.
[17] Z. Kozareva, “Bootstrapping named entity recognition with automatically generated gazetteer lists,” EACL, pp.15–21, 2006.
[18] D. McClosky, E. Charniak, and M. Johnson, “Effective self-training for parsing,” NAACL-HLT, pp.152–159, 2006.
[19] A. Carlson, J. Betteridge, R.C. Wang, E.R. Hruschka Jr., and T.M. Mitchell, “Coupled semi-supervised learning for information extraction,” WSDM, pp.101–110, 2010.
[20] G.J. Lindstone, “Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities,” Transactions of the Faculty of Actuaries, vol.8, pp.182–192, 1920.
[21] F. Jelinek and R.L. Mercer, “Interpolated estimation of Markov source parameters from sparse data,” Workshop on Pattern Recognition in Practice, pp.21–23, 1980.
[22] S.M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” IEEE Trans. Acoust., Speech, Signal Process., vol.35, no.5, pp.400–401, 1987.
[23] I.H. Witten and T.C. Bell, “The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression,” IEEE Trans. Inf. Theory, vol.37, no.4, pp.1085–1094, 1991.
[24] H. Ney and U. Essen, “On smoothing techniques for bigram-based natural language modelling,” ICASSP, vol.2, pp.825–828, 1991.
[25] H. Ney, U. Essen, and R. Kneser, “On structuring probabilistic dependences in stochastic language modeling,” Computer Speech and Language, vol.8, no.1, pp.1–38, 1994.
[26] D.J.C. MacKay and L.C.B. Peto, “A hierarchial Dirichlet language model,” Natural Language Engineering, vol.1, no.3, pp.289–308, 1995.
[27] W.A. Gale and G. Sampson, “Good-Turing frequency estimation without tears,” J. Quantitative Linguistics, vol.2, no.3, pp.217–237, 1995.
[28] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” ICASSP, vol.1, pp.181–184, 1995.

（平成 30 年 6 月 23 日受付，10 月 29 日再受付，
31 年 1 月 11 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．2018 同大学院工学研究科情報・知能工学専攻博士前期課程修了．現在，同博士後期課程在籍．主として，自然言語処理，データマイニングに関する研究に従事．言語処理学会学生会員．

川上　賢十

2017 豊橋技術科学大学情報・知能工学課程卒業．現在，同大学院工学研究科情報・知能工学専攻博士前期課程在籍．主として，自然言語処理，データマイニングに関する研究に従事．

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．

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．