DEIM Forum 2022 K31-1(day3 p39)

閾値近傍の低頻度データに対する保守的な尤度比推定法の提案

日下部友飛† 菊地 真人† 大囿 忠親†

† 名古屋工業大学大学院 情報工学専攻 〒 466–8555 愛知県名古屋市昭和区御器所町
E-mail: kusayu@ozlab.org, {kikuchi,ozono}@nitech.ac.jp

あらまし 本稿では，事象の観測頻度を用いた尤度比推定に取り組む．尤度比の素朴な推定法として，尤度比をなす確率分布を相対頻度で推定し，その比を取る方法がある．しかしこの方法では，低頻度の事象に対する尤度比を過大推定する問題がある．この問題の回避策として，閾値以下の頻度に対する推定値をゼロとする方法がある．この方法では一部の尤度比が計算不要となり，尤度比推定による実用タスクを効率化できる．しかし，閾値近傍の低頻度に対する推定値を，依然として高く見積もってしまう．そこで，閾値近傍の低頻度に対して尤度比を低め（保守的）に見積もる推定法を提案する．尤度比推定に基づくテキストからの文脈予測タスクにおいて，閾値を設けつつ保守的な推定を行うことで，効率性を保ちながら予測精度が向上することを報告する．

キーワード 尤度比，保守的な推定，閾値，低頻度，効率化

1 はじめに

尤度比は統計検定 [1] や二値分類 [2] でよく用いられる統計量であり，その推定法は尤度比を利用するアプリケーションの有用性に大きく影響する．自然言語処理では，テキスト中の文字や単語といった離散的な要素から観測頻度を数え上げ，その頻度に基づいて尤度比を推定することがある [3], [4]．言語資源が含む要素は限定され，その頻度分布はべき乗則に従うため，低頻度の要素が多数を占める．このような状況において，尤度比の素朴な推定法は，低頻度から尤度比を求める際に推定値を不当に高く見積もってしまう問題がある．

上記の問題について，次式で与えられる尤度比 $r(x)$ の推定を例に説明する．

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

ここで，要素 x をテキスト中の文字や単語といった離散要素とする．尤度比の素朴な推定量 $r_{\mathrm{MLE}}(x)$ は次式のように，それぞれの確率分布を相対頻度 $\widehat{p}_*(x)$[1]で求めて比を取る方法である．

$$\widehat{p}_*(x) = \frac{f_*(x)}{n_*}, \quad r_{\mathrm{MLE}}(x) = \frac{\widehat{p}_{\mathrm{nu}}(x)}{\widehat{p}_{\mathrm{de}}(x)}$$

ただし $* \in \{\mathrm{de}, \mathrm{nu}\}$ であり，de は尤度比の分母，nu は分子を表す添え字である．$f_*(x)$ は密度 $p_*(x)$ に従う確率分布からサンプリングされた要素 x の観測頻度であり，$n_* = \sum_x f_*(x)$ である．要素 x_{A} から x_{D} について，頻度に基づく尤度比推定の例を表 1 に示す．要素 x_{A} と x_{B} に着目すると，両者の観測頻度は大きく異なることが分かる．x_{A} は $f_{\mathrm{nu}}(x_{\mathrm{A}}) = 200$ と十分な出現が観測される一方，x_{B} は $f_{\mathrm{nu}}(x_{\mathrm{B}}) = 2$ であり偶然の出現かもしれない．この場合，x_{B} の推定値に対する “信頼性” は，x_{A} の推定値に対する信頼性よりも低いと考えられる．しかし，$r_{\mathrm{MLE}}(x_{\mathrm{A}})$ と $r_{\mathrm{MLE}}(x_{\mathrm{B}})$ は 40 と等しく大きい値になっており，

1：確率分布を多項分布でモデリングした場合の最尤推定量である．

表 1: 尤度比の推定例（閾値 θ_{th}，θ_{L1} は共に 2 とした）．

x	観測頻度				$r_{\mathrm{MLE}}(x)$	$r_{\mathrm{th}}(x)$	$r_{\mathrm{L1}}(x)$
	n_{de}	$f_{\mathrm{de}}(x)$	n_{nu}	$f_{\mathrm{nu}}(x)$			
x_{A}	10^7	5,000	10^4	200	40	40	39.6
x_{B}	10^7	50	10^4	2	40	0	0
x_{C}	10^7	5,000	10^4	300	60	60	59.6
x_{D}	10^7	50	10^4	3	60	60	20

頻度に基づく信頼性を推定値に反映できていない．同様の問題が x_{C} と x_{D} の比較でも確認できる．また，低頻度から推定される $r_{\mathrm{MLE}}(x)$ は変動が大きく不安定である．x_{B} と x_{D} に着目すると，$f_{\mathrm{nu}}(x_{\mathrm{B}}) = 2$，$f_{\mathrm{nu}}(x_{\mathrm{D}}) = 3$ とその差が 1 しかないが，$r_{\mathrm{MLE}}(x_{\mathrm{B}})$ と $r_{\mathrm{MLE}}(x_{\mathrm{D}})$ はそれぞれ 40，60 と大きく異なる．以上のように $r_{\mathrm{MLE}}(x)$ は，頻度の変化による影響を受けやすく，低頻度から推定される尤度比を不当に高く見積もってしまう．

尤度比の過大推定を回避する手段の一つとして，$f_{\mathrm{nu}}(x)$ に閾値を設けることが挙げられる．その推定量は

$$r_{\mathrm{th}}(x) = \begin{cases} r_{\mathrm{MLE}}(x) & \text{if } f_{\mathrm{nu}}(x) > \theta_{\mathrm{th}} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

と定義される．この推定量では，$f_{\mathrm{nu}}(x)$ の頻度があらかじめ定めた閾値 θ_{th} $(\geqq 0)$ より高い場合は $r_{\mathrm{MLE}}(x)$ として推定を行い，閾値以下の場合は他の頻度によらず推定値をゼロにする．これにより，$f_{\mathrm{nu}}(x)$ が閾値以下の低頻度に対して便宜上，過大推定を防ぐ．また，この推定量を用いると閾値以下の頻度に対する計算が不要となるため，尤度比を利用する実用タスクの効率化が期待できる．表 1 中の $r_{\mathrm{th}}(x)$ は θ_{th} を 2 としたときの推定値である．x_{B} に着目すると，$f_{\mathrm{nu}}(x_{\mathrm{B}})$ が 2 であるため，推定値がゼロとなり過大推定が回避される．しかし x_{C} と x_{D} に着目すると，両者の観測頻度は大きく異なるが，$r_{\mathrm{th}}(x_{\mathrm{C}})$ と $r_{\mathrm{th}}(x_{\mathrm{D}})$ は 60 と等しく大きい値になってしまう．これは x_{D} の頻度 $f_{\mathrm{nu}}(x_{\mathrm{D}})$ が 3 であり，$\theta_{\mathrm{th}} = 2$ をわずかに上回るためである．このように，$r_{\mathrm{th}}(x)$ は閾値近傍の低頻度から推定される尤

度比を不当に高く見積もってしまう．また，頻度が閾値以下になると推定値が突然ゼロになり，閾値の前後で推定値が大きく変動するのは値の推移として不自然である．

本稿では，閾値近傍の低頻度から計算される推定値を低めに見積もる推定法を提案する．以降，推定値を低めに見積もることを “保守的な推定” と呼称する．この推定量は，L1 正則化付きの二乗誤差を最小化する最適化の枠組みに基づいて導出され，

$$r_{\mathrm{L1}}(x) = \left(\frac{f_{\mathrm{de}}(x)}{n_{\mathrm{de}}}\right)^{-1} \frac{\max(f_{\mathrm{nu}}(x) - \theta_{\mathrm{L1}},\ 0)}{n_{\mathrm{nu}}}$$

と定義される．ここで，θ_{L1} ($\geqq 0$) は L1 正則化に由来する閾値である．$r_{\mathrm{L1}}(x)$ では，頻度 $f_{\mathrm{nu}}(x)$ が θ_{L1} 以下の場合，$r_{\mathrm{th}}(x)$ と同様に他の頻度によらず推定値をゼロとする．一方で $f_{\mathrm{nu}}(x)$ が閾値よりも高い場合，$f_{\mathrm{nu}}(x)$ を θ_{L1} で減算し，推定値を保守的に見積もる効果を生む．表 1 中の $r_{\mathrm{L1}}(x)$ は $\theta_{\mathrm{L1}} = 2$ としたときの推定値である．まず，閾値と同じ頻度から推定される $r_{\mathrm{L1}}(x_{\mathrm{B}})$ はゼロになっている．次に，高頻度から推定される $r_{\mathrm{L1}}(x_{\mathrm{A}})$ と $r_{\mathrm{L1}}(x_{\mathrm{C}})$ は，他の推定量による推定値とほぼ変わらない，40, 60 に近い値となっている（それぞれ 39.6，59.6）．それに対し，閾値をわずかに上回る低頻度から推定される $r_{\mathrm{L1}}(x_{\mathrm{D}})$ は，20 と保守的に見積もられることが分かる．したがって，提案する推定量 $r_{\mathrm{L1}}(x)$ は閾値を設ける利点を保持し，さらに頻度に応じた保守的な推定も行い，$r_{\mathrm{MLE}}(x)$ と $r_{\mathrm{th}}(x)$ が抱える過大推定の問題を緩和できる．

実験では，固有表現の左に現れる単語バイグラムを尤度比推定によりコーパスから予測する．そして $r_{\mathrm{MLE}}(x)$ および $r_{\mathrm{th}}(x)$ との比較により，保守的な推定が予測精度の向上に寄与することを示す．また閾値の有効性を検証するため，予測に要する計算時間とメモリ使用量も測定し，提案する推定量を用いると効率的な予測ができることも示す．

2 関 連 研 究

尤度比の素朴な推定法として，尤度比を構成する個々の確率分布を相対頻度で求め，その比を取ることが挙げられる．しかし 1 節で述べたように，この方法では推定に用いる頻度に依存して，推定値の変動が大きく安定しない．加えて尤度比を過大推定する場合もある．そこで，確率分布の推定を介さずに尤度比を推定する “直接推定法” が提案された [5], [6], [7], [8]．これまでに様々な直接推定法が提案されているが，これらの推定法は連続的な標本空間で定義される尤度比を推定の対象とし，標本の要素としても連続値を想定している．そこで菊地ら [9] は，最小二乗法に基づく直接推定法 unconstrained Least-Squares Importance Fitting（uLSIF）[8] で使用する基底関数を変更し，uLSIF を離散的な標本空間にも適用可能にした．この推定量は，最適化で導入される L2 正則化の作用により，頻度に閾値を導入せずに保守的な推定を行う．我々の推定量も菊地らの基底関数を用いて同様に定式化されるが，正則化の手法が L2 正則化から L1 正則化になる．この変更によって頻度に閾値を導入する．そして，閾値より大きい頻度の要素にのみ，頻度に基づく保守的な尤度比推定を行う．菊地らの推定量 [9] は全頻度帯の要素に対し，有効な尤度比推定を行うことを目的とする．それに対し，我々の推定量は尤度比を利用する実用タスクの効率化を目的としている．

頻度に閾値を設ける統計量推定のアプローチは単純で，低頻度による過大推定を防ぐだけでなく，統計量を利用する実用タスクの効率化も期待できる．ゆえに，尤度比 [10] 以外の統計量推定（例えば，条件付き確率推定 [11]）にもよく利用されてきた．しかし，既存研究のほとんどは閾値を設ける以上の注意を払っておらず，閾値近傍の頻度による過大推定という問題点には注目していない．青葉ら [12] は，閾値近傍の低頻度から推定される条件付き確率を保守的に見積もる手法を提案した．青葉らの推定量は，導出過程および作用が我々の推定量とよく似ているが，条件付き確率の推定量であることが異なる．条件付き確率は確率の比で表現でき，尤度比の一種ともみなせる．ゆえに本研究の成果は，青葉らの成果を尤度比全般の推定へと拡張したものと解釈することもできる．

3 提 案 手 法

閾値近傍の低頻度に対する尤度比を，保守的に見積もる推定法について述べる．まず，提案する推定量の定式化を説明する．そして，定式化した推定量の数値的な振る舞い分析を行い，閾値を設けた保守的な推定の作用を説明する．

3.1 定 式 化

尤度比推定の問題設定を述べる．あるデータに含まれる要素 x の集合を $D \subset \mathbb{U}$ とする．ここで，x は文字や単語といった離散要素である．$\mathbb{U}$ は存在しうる v 種類の全要素からなる集合であり，情報理論では有限アルファベットとも呼ばれる．いま，密度 $p_{\mathrm{de}}(x)$ を持つ確率分布に従う i.i.d. 標本と，密度 $p_{\mathrm{nu}}(x)$ を持つ確率分布に従う i.i.d. 標本

$$\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{de}}(x), \quad \{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{nu}}(x)$$

を得たとする．これまでの先行研究にならい，密度 $p_{\mathrm{de}}(x)$ が次の条件を満たすと仮定する．

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

この仮定により，全ての x に対して尤度比の定義が可能になる．本節では，標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ と $\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ から尤度比

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

を確率分布の推定を経由せずに直接推定する．なお de は尤度比の分母，nu は分子を表す添え字である．

最小二乗法に基づく直接推定法 unconstrained Least-Squares Importance Fitting（uLSIF）[8] では，$r(x)$ を線形和

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

でモデル化する．$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_b)^{\mathrm{T}}$ は標本から学習されるパラメータ，$\{\varphi_l\}_{l=1}^{b}$ は非負値を取る基底関数である．本来の

uLSIF は基底関数の定義中にガウスカーネルを用いた．しかし，離散的な標本空間で定義される尤度比を推定対象とする場合，この基底関数は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，菊地ら [9] によって提案された，要素の種類ごとに異なる基底関数 $\{\delta_l\}_{l=1}^{v}$

$$\delta_l(x) = \begin{cases} 1 & \text{if } x = x_{(l)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

を代用する．添え字 l は存在しうる v 種類の要素から特定の要素を指定する．すなわち，$x_{(l)}$ は v 種類の要素のうち，l 種類目の要素を表す．式 (2) の基底関数を使用すると，$x_{(m)},\ m = 1, 2, \ldots, v$ に対する推定モデルは

$$\widehat{r}(x_{(m)}) = \sum_{l=1}^{v} \beta_l \delta_l(x_{(m)}) = \beta_m \tag{3}$$

となる．uLSIF では，推定モデル $\widehat{r}(x_{(m)})$ と真の尤度比 $r(x_{(m)})$ の二乗誤差を最小化するパラメータ $\boldsymbol{\beta}$ を求める．本来の uLSIF では，過学習を防ぐため二乗誤差の最小化に L2 正則化を導入する．対して我々は，一部の頻度に対する計算を省くために，正則化の手法を L2 正則化から L1 正則化へと置き換える．その最適化問題は次式で与えられる [2]．

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^v} \left[\frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \widehat{\boldsymbol{H}} \boldsymbol{\beta} - \widehat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{\beta} + \lambda_{\mathrm{L1}} \sum_{l=1}^{v} \beta_l \right] \tag{4}$$

$\mathbb{R}^v$ は実 v 次元空間である．上式では $\boldsymbol{\beta}$ に対する正則化のためにペナルティ項 $\lambda_{\mathrm{L1}} \sum_{l=1}^{v} \beta_l$ を導入する．λ_{L1} $(\geqq 0)$ は正則化パラメータである．なお L1 正則化項は本来，パラメータ β_l に対する絶対値の総和 $\sum_{l=1}^{v} |\beta_l|$ となるが，上式では絶対値が外れている．式 (3) に示すように，β_l は尤度比の推定モデルそのものであり，尤度比の非負性により β_l も非負性を持つ．ゆえに β_l の絶対値を外すことができる．$\widehat{\boldsymbol{H}}$ は $v \times v$ 行列であり，その (l, l') 番目の要素 $\widehat{H}_{l,l'}$ は次式で定義される．

$$\begin{aligned} \widehat{H}_{l,l'} &= \frac{1}{n_{\mathrm{de}}} \sum_{i=1}^{n_{\mathrm{de}}} \delta_l(x_i^{\mathrm{de}}) \delta_{l'}(x_i^{\mathrm{de}}) \\ &= \begin{cases} \frac{f_{\mathrm{de}}(x_{(l)})}{n_{\mathrm{de}}} & \text{if } l = l' \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

上式からわかるように $\widehat{\boldsymbol{H}}$ は対角行列になる．また，$\boldsymbol{h}$ は v 次元ベクトルであり，その l 番目の要素 $\widehat{h}_l$ は次式で定義される．

$$\widehat{h}_l = \frac{1}{n_{\mathrm{nu}}} \sum_{j=1}^{n_{\mathrm{nu}}} \delta_l(x_j^{\mathrm{nu}}) = \frac{f_{\mathrm{nu}}(x_{(l)})}{n_{\mathrm{nu}}}$$

ここで $f_*(x_{(l)})$ は，密度 $p_*(x)$，$* \in \{\mathrm{de}, \mathrm{nu}\}$ を持つ確率分布からサンプリングされた $x_{(l)}$ の観測頻度である．式 (4) の目的関数において 1 項目と 2 項目は二乗誤差に由来する項であり，式 (4) は拘束無し二次計画問題である．ベクトルや行列の成分を用いると，1 項目と 2 項目はそれぞれ

$$\frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \widehat{\boldsymbol{H}} \boldsymbol{\beta} = \frac{1}{2} \sum_{l=1}^{v} \frac{f_{\mathrm{de}}(x_{(l)})}{n_{\mathrm{de}}} \beta_l^2 \tag{5}$$

$$-\widehat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{\beta} = -\sum_{l=1}^{v} \frac{f_{\mathrm{nu}}(x_{(l)})}{n_{\mathrm{nu}}} \beta_l \tag{6}$$

と表される．式 (4) の目的関数を β_m で偏微分してゼロと置く．

$$\frac{\partial}{\partial \beta_m} \left(\frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \widehat{\boldsymbol{H}} \boldsymbol{\beta} - \widehat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{\beta} + \lambda_{\mathrm{L1}} \sum_{l=1}^{v} \beta_l \right) = 0$$

そして上式に式 (5)，(6) を代入し β_m について解くと，二乗誤差を最小化するパラメータ

$$\begin{aligned} \widehat{r}(x_{(m)}) &= \widetilde{\beta}_m(\lambda_{\mathrm{L1}}) \\ &= \left(\frac{f_{\mathrm{de}}(x_{(m)})}{n_{\mathrm{de}}} \right)^{-1} \left(\frac{f_{\mathrm{nu}}(x_{(m)})}{n_{\mathrm{nu}}} - \lambda_{\mathrm{L1}} \right) \end{aligned}$$

が得られる．式 (3) からも分かるように，$\widetilde{\beta}_m(\lambda_{\mathrm{L1}})$ が要素 $x_{(m)}$ に対する尤度比の推定量 $\widehat{r}(x_{(m)})$ となる．なお上式では，λ_{L1} が $\frac{f_{\mathrm{nu}}(x_{(m)})}{n_{\mathrm{nu}}}$ よりも大きいとき，$\widehat{r}(x_{(m)})$ が負になる．しかしながら尤度比の真値は必ず非負である．uLSIF では，式 (4) の解が負の値を取る場合，負の値をゼロに丸める近似をする．そこで上式が負になる場合も同様に，$\widehat{r}(x_{(m)})$ をゼロに補正する．以上より，提案する推定量は

$$r_{\mathrm{L1}}(x_{(m)}) = \left(\frac{f_{\mathrm{de}}(x_{(m)})}{n_{\mathrm{de}}} \right)^{-1} \max \left(\frac{f_{\mathrm{nu}}(x_{(m)})}{n_{\mathrm{nu}}} - \lambda_{\mathrm{L1}},\ 0 \right)$$

と定義される．上式では，尤度比の分子 $\frac{f_{\mathrm{nu}}(x_{(m)})}{n_{\mathrm{nu}}}$ が正則化パラメータ λ_{L1} 以下の場合，他の頻度によらず推定値をゼロとする．一方で $\frac{f_{\mathrm{nu}}(x_{(m)})}{n_{\mathrm{nu}}}$ が λ_{L1} よりも大きい場合，分子の値を λ_{L1} で減算して，推定値を保守的に見積もる．

正則化パラメータ λ_{L1} の値はゼロ以上の実数として自由に設定できる．本稿では，我々の推定量 $r_{\mathrm{L1}}(x_{(m)})$ と単に閾値を設ける推定量との対等な比較を行う目的で，$\lambda_{\mathrm{L1}} = \frac{\theta_{\mathrm{L1}}}{n_{\mathrm{nu}}}$ と定義する．この定義により，我々の推定量は頻度に閾値を設ける形式となり，4 節で述べる実験において，閾値近傍の頻度に対する保守的な推定の有効性を検証できる．このとき，$r_{\mathrm{L1}}(x_{(m)})$ は

$$r_{\mathrm{L1}}(x_{(m)}) = \left(\frac{f_{\mathrm{de}}(x_{(m)})}{n_{\mathrm{de}}} \right)^{-1} \frac{\max(f_{\mathrm{nu}}(x_{(m)}) - \theta_{\mathrm{L1}},\ 0)}{n_{\mathrm{nu}}} \tag{7}$$

と置き換えられる．ここで，θ_{L1} $(\geqq 0)$ は頻度 $f_{\mathrm{nu}}(x_{(m)})$ に対する閾値である．この式では，$f_{\mathrm{nu}}(x_{(m)})$ が小さいほど，尤度比の推定値がより保守的に見積もられる．以上より，提案する推定量は，閾値近傍の低頻度に対して尤度比の過大推定を抑制する．加えて，閾値以下の頻度に対する計算を省き，尤度比を用いる実用タスクの効率化を図る．なお θ_{L1} がゼロのとき，上式は $r(x_{(m)})$ の分母と分子の確率分布をそれぞれ相対頻度で推定し，その比を取った推定量に等しい．

3.2 推定量の数値的な振る舞い分析

提案手法によって計算された推定値が，観測頻度に応じてどう変化するのかを分析したい．そのために，要素 x に対するそ

2：式 (4) の導出過程は uLSIF の原論文 [8] を参照のこと．

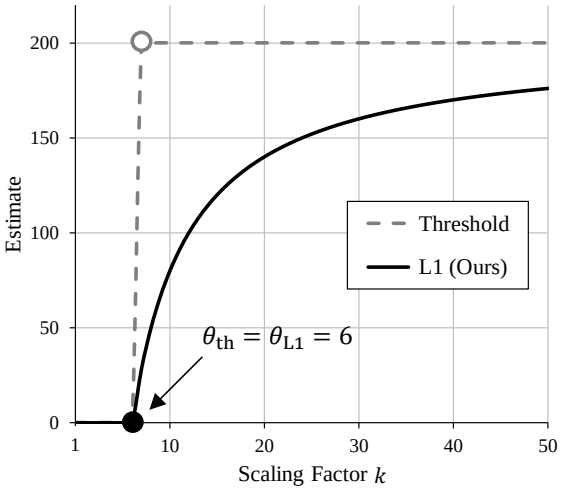


図 1: 観測頻度による推定値の変化．$r_{\mathrm{MLE}}(x)$ が頻度の変化によらず，一定（200）になる状況を仮定している．

れぞれの頻度を次のように定義する．

$$n_{\mathrm{de}} = 10^7,\ f_{\mathrm{de}}(x) = 5 \times k,$$
$$n_{\mathrm{nu}} = 10^4,\ f_{\mathrm{nu}}(x) = k$$

k は任意の自然数を取るスケーリング係数であり，これを調節して頻度を変化させる．これらの頻度をもとに尤度比 $r(x) = \frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}$ を推定する．k が大きいときは $f_{\mathrm{de}}(x)$ と $f_{\mathrm{nu}}(x)$ が高くなり，高頻度から $r(x)$ が推定される．逆に k が小さいときは，低頻度から $r(x)$ が推定される．

式 (1) に示した閾値による単純な推定量 $r_{\mathrm{th}}(x)$，式 (7) に示した提案する推定量 $r_{\mathrm{L1}}(x)$ の振る舞いを比較する．なお，確率分布 $p_*(x)$ を相対頻度 $\widehat{p}_*(x) = \frac{f_*(x)}{n_*}$，$* \in \{\mathrm{de}, \mathrm{nu}\}$ で求め，その比を取った推定値は k によらず

$$r_{\mathrm{MLE}}(x) = \frac{\widehat{p}_{\mathrm{nu}}(x)}{\widehat{p}_{\mathrm{de}}(x)} = \left(\frac{5}{10^7}\right)^{-1}\left(\frac{1}{10^4}\right) = 200$$

と一定になる．閾値を $\theta_{\mathrm{th}} = \theta_{\mathrm{L1}} = 6$ とした場合の $r_{\mathrm{th}}(x)$ と $r_{\mathrm{L1}}(x)$ の振る舞いを図 1 に示す．この図は横軸をスケーリング係数 k，縦軸を k に対応した推定値とするグラフである．横軸の値が小さいほど推定値が低頻度から計算され，横軸の値が大きいほど推定値が高頻度から計算される．$r_{\mathrm{th}}(x)$ は $f_{\mathrm{nu}}(x) = 7$ まで 200 と大きな推定値だが，$f_{\mathrm{nu}}(x) = \theta_{\mathrm{th}} = 6$ になると途端にゼロになる．すなわち，$r_{\mathrm{th}}(x)$ は過大推定を防ぐ手段として，$f_{\mathrm{nu}}(x)$ が閾値以下か否かしか考慮できない．その結果，閾値をわずかに上回る $f_{\mathrm{nu}}(x) = 7$ の場合でも，$r_{\mathrm{th}}(x)$ は 200 と大きな推定値になってしまう．対して，$r_{\mathrm{L1}}(x)$ は頻度の低さに応じて推定値を保守的に見積もることができる．結果として，閾値をわずかに上回る $f_{\mathrm{nu}}(x) = 7$ から 10 付近では，$r_{\mathrm{L1}}(x)$ は $r_{\mathrm{th}}(x) = 200$ よりもはるかに低いことが分かる．

4 評価実験

固有表現の左に出現する文脈を尤度比によりコーパスから予測する．この実験により，提案手法の有効性を文脈の予測精度

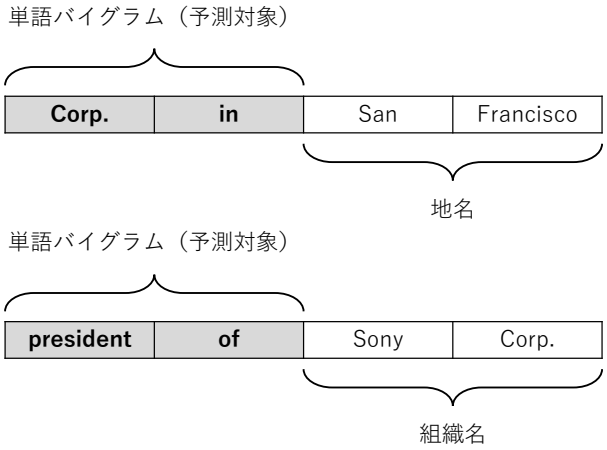


図 2: 固有表現の左にある単語バイグラムの例．

表 2: 使用する実験データ．左列から順に，データ全体に含まれるバイグラムの種類数と総頻度，それぞれの固有表現の左に出現するバイグラムの種類数と総頻度を表す．

データ	全体		地名の左		組織名の左	
	種類数	総頻度	種類数	総頻度	種類数	総頻度
訓練	1,468,292	4,002,930	31,294	64,072	44,946	94,737
開発	230,528	401,445	4,318	6,116	6,443	9,946
評価	231,931	403,145	4,164	5,876	6,544	9,857

と予測効率の観点から評価する．図 2 に示すように，使用する固有表現を地名と組織名とし，予測する文脈を地名あるいは組織名の左に位置する単語バイグラムとする．出現文脈を予測する理由は次の二点である．第一に，バイグラムなどの言語要素は種類が豊富だが，低頻度のものが大半を占める．この場合，低頻度の扱い方により，予測の精度や効率が大きく変化する．よって，頻度の扱い方が異なる尤度比推定法の差異が明確になりやすく，提案手法の有効性検証に適する．第二に，出現文脈は正解が一意に定まるため，提案手法を定量評価できる．

4.1 実験環境

実験環境を以下に示す．

- OS：Windows 10 Home
- プロセッサ：AMD Ryzen 7 Extreme Edition @ 1.80GHz
- メモリ：16.0GB
- Python：3.6.10

4.2 実験データと実験条件

ウォール・ストリート・ジャーナルコーパス [3] の 1987 年版を用い，次の手順で実験データを作成する．コーパスからランダムに 12,000 記事をサンプリングする．Stanford Named Entity Recognizer（Stanford NER）[4] [13] を用いて 12,000 記事に固有表現タグ（地名および組織名）を付与する．10,000 記事を訓練用に，残りの 2,000 記事を開発用と評価用にそれぞれ 1,000 記事ずつ割り振る．訓練用の記事を単語バイグラム単位に分割

3：https://catalog.ldc.upenn.edu/LDC2000T43
4：https://nlp.stanford.edu/software/CRF-NER.html

し，固有表現の左，10,000 記事全体での出現頻度をバイグラムの種類ごとに計数する．このようにして，頻度情報をまとめたものを訓練データとする．また，開発用と評価用の記事が含む全バイグラムの集合を，それぞれ開発データ，評価データとする．なお開発用，評価用の記事に付与された固有表現タグは，正誤判定にのみ利用する．実験データの情報を表 2 に示す．バイグラムの種類数と総頻度は 1.5〜2 倍程度しか変わらず，低頻度のバイグラムが多くを占めることが分かる．

実験条件として固有表現の種類を指定する．固有表現は地名と組織名の二種類があり，あらかじめどちらか一方を選択する．

4.3 実験手順

まず，訓練データから尤度比の推定に必要な頻度情報を格納する．ただし閾値のある手法では，閾値以下の頻度を持つバイグラムの情報は推定に不要なため格納しない．次に格納した頻度情報を用い，評価データにある全バイグラムに対して尤度比

$$r(x) = \frac{p(x|c_{\mathrm{NE}})}{p(x)} \tag{8}$$

を推定する．x はバイグラムであり，c_{NE} は固有表現の左バイグラムに付与されるクラスラベルである．$p(x|c_{\mathrm{NE}})$ は固有表現の左における x の出現確率であり，$p(x)$ は訓練用の記事全体における x の出現確率である．

尤度比の推定法ごとに，バイグラムを推定値の降順にソートし，その上位 4,000 件を正誤判定する．評価データでバイグラムが固有表現の左に一度でも出現すれば正解，それ以外は不正解とする．正誤判定の結果を用い，4.4 節に示す尤度比の推定法ごとにランク—再現率曲線を描く．この曲線は，横軸をバイグラムのランク（順位），縦軸をそのランクでの再現率としたグラフ上に描かれる．あるランクで最高の再現率を持つ推定法が，そのランクでの最良の手法となる．この曲線では，曲線上のある点とグラフの原点を結んだ直線の傾きが，そのランクでの適合率に比例する．再現率と適合率はそれぞれ

$$\text{再現率} = \frac{|\{x \mid x \in R \cap X\}|}{|R|},$$
$$\text{適合率} = \frac{|\{x \mid x \in R \cap X\}|}{|X|}$$

と定義される．X は上位 4,000 件のバイグラムの集合，R は評価データにおいて固有表現の左に出現するバイグラムの集合，すなわち正解集合を意味する．

また，閾値を設けることによる予測タスクの効率化を確認するため，計算時間とメモリ使用量を計測する．計算時間とは，訓練に使用するバイグラムの頻度を格納し，評価データに含まれる全バイグラムに対して尤度比を推定し終わるまでの時間とする．なお計算時間は，上述の実験手順を 10 回繰り返した際の平均値として算出する．メモリ使用量は，訓練に使用するバイグラムの頻度を全て格納するのに要するメモリ使用量とする．

4.4 比較手法

尤度比の推定法として以下の 4 手法を比較する．推定する尤度比を $r(x) = \frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}$ として各推定法を説明する．なお式 (8) に示すように本実験では，$p_{\mathrm{de}}(x)$ が $p(x)$，$p_{\mathrm{nu}}(x)$ が $p(x|c_{\mathrm{NE}})$ にそれぞれ対応している．手法 1 と手法 2 は閾値を設けない手法，手法 3 と手法 4 は閾値を設ける手法である．

手法 1：Baseline　確率分布 $p_*(x)$ を相対頻度 $\widehat{p}_*(x) = \frac{f_*(x)}{n_*}$，$* \in \{\mathrm{de}, \mathrm{nu}\}$ で求め，その比を取った推定量 $r_{\mathrm{MLE}}(x)$ を Baseline とする．$r_{\mathrm{MLE}}(x)$ は

$$r_{\mathrm{MLE}}(x) = \frac{\widehat{p}_{\mathrm{nu}}(x)}{\widehat{p}_{\mathrm{de}}(x)}$$

と定義される．なお，評価データにあるバイグラムが訓練データにない場合，$r_{\mathrm{MLE}}(x)$ はゼロ除算のため推定値を計算できなくなる．この場合は推定値をゼロとして対処した．Baseline を用いると単純に尤度比推定できるが，低頻度から求まる推定値を不当に高く見積もってしまう．

手法 2：L2　L2 正則化により，頻度に応じて推定値を保守的に見積もる．$r_{\mathrm{L2}}(x)$ は

$$r_{\mathrm{L2}}(x) = \left(\frac{f_{\mathrm{de}}(x)}{n_{\mathrm{de}}} + \lambda_{\mathrm{L2}}\right)^{-1} \frac{f_{\mathrm{nu}}(x)}{n_{\mathrm{nu}}}$$

と定義される．この推定量は，L2 正則化付きの二乗誤差を最小化する枠組みにより導出される．λ_{L2} $(\geqq 0)$ は正則化パラメータであり，これが推定値を保守的に見積もる効果を生む．ただし，$r_{\mathrm{L2}}(x)$ は頻度に閾値を設けないため，本実験における文脈予測タスクの効率化は期待できない．

手法 3：Threshold　$r_{\mathrm{th}}(x)$ は閾値 θ_{th} を設け，それよりも大きい頻度のバイグラムにのみ，頻度に基づく尤度比推定を行う．式 (1) にも示したように $r_{\mathrm{th}}(x)$ は

$$r_{\mathrm{th}}(x) = \begin{cases} r_{\mathrm{MLE}}(x) & \text{if } f_{\mathrm{nu}}(x) > \theta_{\mathrm{th}} \\ 0 & \text{otherwise} \end{cases}$$

と定義される．この推定法は θ_{th} を設けることで便宜上，低頻度による尤度比の過大推定を防ぐ．加えて，一部の尤度比を頻度から推定する必要がないため，文脈予測タスクの効率化も期待できる．しかし $r_{\mathrm{th}}(x)$ には，θ_{th} 近傍の低頻度から推定される尤度比を不当に高く見積もる問題が残っており，過大推定への対処が不十分と考える．

手法 4：L1 (Ours)　閾値より大きい頻度のバイグラムに対する尤度比を，$r_{\mathrm{MLE}}(x)$ よりも保守的に見積もる．式 (7) にも示したように，我々の推定量 $r_{\mathrm{L1}}(x)$ は

$$r_{\mathrm{L1}}(x) = \left(\frac{f_{\mathrm{de}}(x)}{n_{\mathrm{de}}}\right)^{-1} \frac{\max(f_{\mathrm{nu}}(x) - \theta_{\mathrm{L1}},\ 0)}{n_{\mathrm{nu}}}$$

と定義される．この推定量は，L1 正則化付きの二乗誤差を最小化する枠組みにより導出される．$f_{\mathrm{nu}}(x)$ を閾値 θ_{L1} で減算して推定値を保守的に見積もり，θ_{L1} 近傍の低頻度に対する尤度比の過大推定を抑制できる．

手法 2〜4 はそれぞれハイパーパラメータ λ_{L2}，θ_{th}，θ_{L1} を持つ．手法 2 と 3 が持つパラメータの値は次の手順で決定した．まず開発データを評価データとみなし，各手法についてパラメータの値ごとにランク—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の値を最適値として採用し

た．手法 2 では，λ_{L2} を $10^{-9}, 10^{-8}, \ldots, 10^{-1}$ と変化させ，最適値は地名，組織名ともに 10^{-2} となった．手法 3 では，θ_{th} を $9, 8, \ldots, 1$ と変化させ，最適値は地名，組織名ともに 2 となった．提案手法（手法 4）では，閾値による推定法（手法 3）との対等な比較を行うため，θ_{L1} の値を θ_{th} と同様に 2 とした [5]．

4.5 実験結果

ランク—再現率曲線を図 3 に示す．この曲線は，バイグラムを推定値の降順に並べた際のランク（順位）を横軸とし，そのランクでの再現率を縦軸としたグラフ上に描かれる．あるランクで最高の再現率を持つ推定法が，そのランクでの最良の推定法となる．この曲線では，曲線上のある点とグラフの原点を結んだ直線の傾きが，そのランクでの適合率に比例する．Baseline は他の推定法よりもバイグラムの予測性能が劣る．これは，Baseline が低頻度バイグラムの尤度比を過大推定し，高頻度から推定される “確からしい” バイグラムを下位に追いやるためと考える．特に，組織名の左バイグラムは種類が豊富で低頻度のものも多く，図 3(b) に示すように Baseline の性能の低さが際立つ．L2 は低頻度バイグラムの尤度比を保守的に見積もり，高頻度から推定されるバイグラムを上位にできる．そのため，地名と組織名の両方で最高の性能を達成している．閾値を設けた推定法 Threshold および L1（提案手法）は Baseline よりも優れている．これは閾値によって尤度比の過大推定が防止できたためと考える．また L1 は Threshold よりも，わずかに良好なことが確認できる．したがって，閾値近傍における保守的な尤度比推定は，単純に閾値を設けるアプローチよりも有効なことが示唆された．ただし，閾値を設ける 2 手法は閾値以下の低頻度バイグラムに対する尤度比を一律にゼロとする．この作用によって，L2 よりも全体的に性能が低下し，下位になると再現率が向上しなくなる悪影響も確認できた．

計算時間とメモリ使用量の測定結果を表 3 に示す．まず，閾値を設けない手法のグループ（Baseline，L2）と閾値を設ける手法のグループ（Threshold，L1）とを比較する．この比較から，閾値を設けると計算時間は 1/4 程度，メモリ使用量は 1/10 程度の効率化が確認できる．加えて，各手法が訓練時に用いたバイグラムの種類数を表 4 に示す．この表から分かるように，どの手法を見ても，用いたバイグラムの種類数は訓練用の記事が含む全バイグラムの種類数（表 2）よりもはるかに少ない．これは閾値を設けない手法であっても，固有表現の左に現れるバイグラムの頻度情報のみを格納するためである．閾値を設けた手法では閾値が 2 であるから，固有表現の左に 3 回以上現れるバイグラムの頻度情報を格納する．閾値がある場合は閾値がない場合と比較して，バイグラムの種類数が約 1/10 となっており，これはメモリ使用量の測定結果と概ね一致する．Threshold と異なり L1 は，保守的な推定を行う際に頻度を閾値で減算する処理が加わる．そこで次に，閾値を設ける二つの推定法 Threshold と L1 を比較すると，計算時間とメモリ使用量の両方にほぼ差がないことが確認できる．よって，文脈予測タスクを解く効率性の観点からも L1 の有効性が示唆された．

5：念のため，提案手法でも曲線下面積が最大となるパラメータの値を調査した．その結果，θ_{L1} の最適値は θ_{th} と同様に 2 であることが分かった．

5 おわりに

本稿では，閾値近傍の低頻度に対する尤度比の保守的な推定法を提案した．この手法では，頻度に閾値を設けて推定を効率化し，閾値をわずかに上回る低頻度に対する尤度比の過大推定も抑制できる．提案する推定量は二乗誤差を最小化する理論的な枠組みで導出され，L1 正則化の作用で保守的な推定を実現する．固有表現の出現文脈を尤度比でコーパスから予測する実験を行い，予測の精度と効率の観点から提案手法の有効性を検証した．その結果，閾値近傍の低頻度に対して保守的な推定を行うことで，単純に閾値を設けるよりも予測精度が改善することが示唆された．また，閾値を設けることで計算時間は 1/4，メモリ使用量は 1/10 の効率化を確認できた．

ただし本稿の実験では，提案手法と単純に閾値を設ける推定法との性能差が小さい．またこの実験は，頻度に閾値を設けずとも容易に完了できる．そこで頻度の閾値が不可欠で，二手法間の差が明瞭になるタスクによって提案手法の有効性を検証したい．候補のタスクとして，組み合わせの列挙が必要なパターンマイニングが考えられる．このタスクでは，パターンを全列挙すると組み合わせ爆発を起こすため，頻度に閾値を設けて見込みのないパターンを効率的に枝刈りする必要がある．また，提案手法は低頻度を有効に扱えるため，単純に閾値を設けるよりも幅広い頻度を推定に利用できる可能性がある．このような有効性も今後確認していきたい．

謝　辞

本研究の一部は JSPS 科研費 JP19K12266 の助成を受けたものです．

文　献

[1] S. Glover and P. Dixon. Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, Vol. 11, No. 5, pp. 791–806, 2004.

[2] 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. 電子情報通信学会技術研究報告, Vol. 114, No. 502, pp. 55–62, 2015.

[3] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, Vol. 19, No. 1, pp. 61–74, 1993.

[4] C. D. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[5] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, pp. 601–608, 2007.

[6] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proc. ICML'07*, pp. 81–88, 2007.

[7] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, pp.

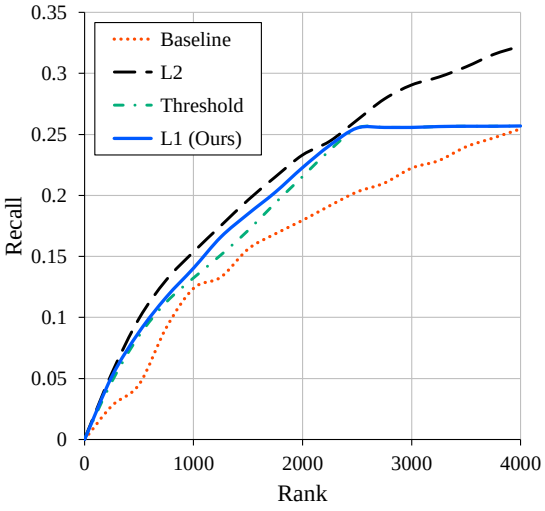
表 3: 計算時間とメモリ使用量．計算時間は 10 回測定した平均値である．

(a) 地名

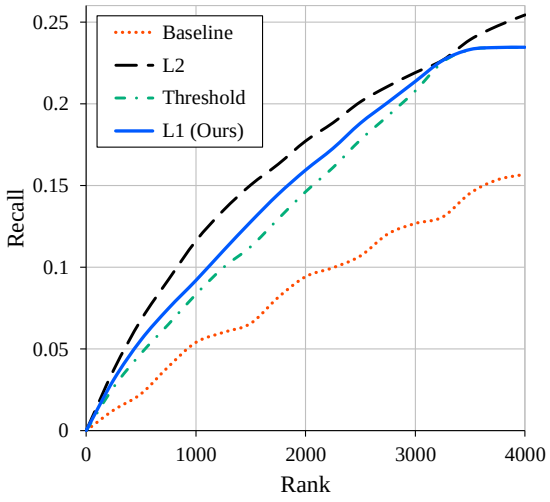
手法	閾値	時間 [sec]	メモリ [KB]
Baseline	なし	3.179	8,180
L2	なし	3.179	8,180
Threshold	2	0.822	854
L1 (Ours)	2	0.823	854

(b) 組織名

手法	閾値	時間 [sec]	メモリ [KB]
Baseline	なし	3.518	13,292
L2	なし	3.289	13,292
Threshold	2	0.838	1,015
L1 (Ours)	2	0.836	1,015



(a) 地名



(b) 組織名

図 3: ランク—再現率曲線.

表 4: 訓練に用いたバイグラムの種類数.

手法	閾値	固有表現	
		地名	組織名
Baseline L2	なし	31,294	44,946
Threshold L1 (Ours)	2	3,207	4,097

1433–1440, 2008.

[8] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, Vol. 10, pp. 1391–1445, July 2009.

[9] 菊地真人, 川上賢十, 吉田光男, 梅村恭司. 観測頻度に基づくゆう度比の保守的な直接推定. 電子情報通信学会論文誌 D, Vol. J102-D, No. 4, pp. 289–301, 2019.

[10] 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.

[11] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. VLDB'94*, Vol. 1215, pp. 487–499, 1994.

[12] T. Aoba, M. Kikuchi, M. Yoshida, and K. Umemura. Improving association rule mining for infrequent items using direct importance estimation. In *Proc. ICAICTA'20*, 2020.

[13] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. ACL'05*, pp. 363–370, 2005.