特徴重み付けを用いた低・ゼロ頻度 N-gram に対する尤度比の推定法

菊地　真人†　　吉田　光男†　　梅村　恭司†

† 豊橋技術科学大学情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1–1
E-mail: m143313@edu.tut.ac.jp, yoshida@cs.tut.ac.jp, umemura@tut.jp

あらまし　本稿では，N-gram の頻度情報から尤度比を推定する問題を扱う．近年，低頻度から尤度比の安定した推定値を求める手法が提案されたが，テキスト上で未観測の（つまりゼロ頻度の）N-gram に対しては尤度比が推定できない．その対処法として，N-gram を成す単語や文字といった特徴間に統計的な独立性を仮定し，特徴ごとに推定した尤度比の積を取る方法が考えられる．このアプローチは単純だが，特徴間の独立性は実際には成立しないことが多い．そこで，ナイーブベイズ分類器で独立性仮定の軽減に多用される，特徴重み付け法を前述の対処法に組み合わせることを提案する．尤度比を使用した文脈予測の実験で，提案手法が有効に作用し良好な結果が得られることを報告する．

キーワード　尤度比推定，低頻度，ゼロ頻度，N-gram，特徴重み付け法

1　はじめに

尤度比は統計検定 [1] や二値分類 [2] で多用される尺度であり，その推定法は尤度比を用いる幅広いアプリケーションの有用性に影響を及ぼす．自然言語処理では，テキスト中の単語や文字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]．一般に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．また，言語資源を用いた教師あり学習では，学習に用いた資源中で未観測（すなわちゼロ頻度）の要素が入力として与えられることも多く，そのような要素に対しても何らかの推定値を算出することが要求される．このような状況において，頻度を用いた単純な推定法は有効な推定値を算出できないことが問題となる．

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．

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

いま，$A_1, A_2, \ldots, A_{\mathrm{N}}$ を N 個の特徴変数，離散要素 $x = \langle a_1, a_2, \ldots, a_{\mathrm{N}} \rangle$ をそれぞれの確率密度に従う確率分布からサンプリングされた特徴ベクトルとする．a_k $(1 \leqq k \leqq \mathrm{N})$ は A_k の取る値であり，単語や文字などの離散値とする．このとき，N 個の離散値の連なりである要素 x は N-gram と呼ばれる．素朴な尤度比の推定法は次式のように，それぞれの確率分布を最尤推定 [1]で求めて比を取る方法である．

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

ただし，添え字 $*$ は “de”，“nu” のいずれかを表す．$f_*(x)$ 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．このアプローチは確率分布の推定を介するため，“間接推定法” と呼ばれる．$r_{\mathrm{MLE}}(x)$ による尤度比の推定例を表 1 に示す．まず x_{A} と x_{B} に着目すると，両者の観測頻度が大きく異なるにもかかわらず，推定値 $r_{\mathrm{MLE}}(x_{\mathrm{A}})$ と $r_{\mathrm{MLE}}(x_{\mathrm{B}})$ は共に 20 と高い値を示す．ここで $f_{\mathrm{nu}}(x_{\mathrm{B}}) = 1$ は偶然の出現であることが十分に考えられ，$r_{\mathrm{MLE}}(x_{\mathrm{B}})$ は信用できない．次に，低頻度から計算される $r_{\mathrm{MLE}}(x)$ は変動が大きく不安定である．例えば x_{B} と x_{C} を比較すると，$f_{\mathrm{nu}}(x)$ の差が 1 しかない場合でも，推定値 $r_{\mathrm{MLE}}(x_{\mathrm{B}})$ と $r_{\mathrm{MLE}}(x_{\mathrm{C}})$ は 20, 40 と大きく異なる．以上のケースでは，観測頻度の低さに応じて推定値が調節されるべきである．最後に x_{D} に着目すると，$f_{\mathrm{nu}}(x_{\mathrm{D}}) = 0$ のために他の頻度に依らず，$r_{\mathrm{MLE}}(x_{\mathrm{D}})$ はゼロになる．同様に $f_{\mathrm{de}}(x) = 0$ の場合も，$r_{\mathrm{MLE}}(x)$ が無限大となって，有効な推定値を算出できない．

表 1: 観測頻度に基づいた尤度比の推定例.

N-gram	観測頻度				$r_{\mathrm{MLE}}(x)$	$\widehat{r}(x)$
x	n_{de}	$f_{\mathrm{de}}(x)$	n_{nu}	$f_{\mathrm{nu}}(x)$		
x_{A}	10^7	5,000	10^4	100	20	19.6
x_{B}	10^7	50	10^4	1	20	6.7
x_{C}	10^7	50	10^4	2	40	13.3
x_{D}	10^7	14	10^4	0	0	0

低頻度に起因する問題に対しては，対処法 [5] が提案されており，その推定式は次式で表される．

$$\widehat{r}(x) = \left(\frac{1}{n_{\mathrm{de}}} f_{\mathrm{de}}(x) + \lambda \right)^{-1} \times \frac{1}{n_{\mathrm{nu}}} f_{\mathrm{nu}}(x)$$

この手法は，二乗損失の最小化プロセスにより尤度比を直接推定する unconstrained Least-Squares Importance Fitting（uLSIF）[6] という枠組みに基づいている．uLSIF の枠組みで導入される正則化パラメータ λ $(\geqq 0)$ が，頻度の低さに応じた推定値の調節を可能にし推定量をロバストにする．表 1 の $\widehat{r}(x)$ は $\lambda = 10^{-5}$ としたときの推定量である．十分な頻度から推定された $\widehat{r}(x_{\mathrm{A}})$ は 19.6 となり $r_{\mathrm{MLE}}(x_{\mathrm{A}})$ に近い一方，低頻度から推定された $\widehat{r}(x_{\mathrm{B}})$ と $\widehat{r}(x_{\mathrm{C}})$ はそれぞれ 6.7 と 13.3 となり，頻度に応じて推定量が調節されたことがわかる．この手法は低頻度の要素から計算される尤度比にも，安定した推定値を付与で

1：厳密には，確率分布を多項分布でモデリングした場合の最尤推定である．

きる有効な方法である．しかし，この手法でも $\widehat{r}(x_{\mathrm{D}})$ は依然としてゼロであり，効果的な推定値を算出できていない．前述のように，言語資源が含む要素は低頻度のものが多数を占める．加えて，この状況では資源中に出現しない要素も多いことが予想され，尤度比の実用を想定すると低・ゼロ頻度の両方に対処できる推定法が望ましい．

そこで本稿では，先の推定量 $\widehat{r}(x)$ をベースとして，低・ゼロ頻度 N-gram に有効な尤度比の推定法を提案する．ゼロ頻度に起因する問題を緩和する簡易な方法として，a_k $(1 \leqq k \leqq \mathrm{N})$ の出現に統計的独立を仮定し，a_k に対する尤度比の積を取ることが考えられる．このとき，$r(x)$ は

$$r(x) = \prod_{k=1}^{\mathrm{N}} r(a_k)$$

と表される．上式では言語資源に x が出現せずとも，それを構成する a_k が単独で出現すれば $r(x)$ の計算が可能になる．また $r(a_k)$ の推定に文献 [5] の直接推定法を利用すれば，低頻度の a_k に対しても安定な推定値が得られる．しかし，ここで仮定した独立性は実際に成立しないことが多く，尤度比の推定性能を低下させる一因となる．そこで，独立性仮定を軽減する手段として，ナイーブベイズ分類器で条件付き独立性仮定の軽減に使用される特徴重み付け法 [7] に着目する．この手法は x を分類するクラスと特徴 A_k の二つが与えられれば，重み付けの対象が尤度比であっても応用できる．また $r(a_k)$ を重み付けすることで，a_k の頻度をそのまま利用でき，尤度比の推定において独立性仮定の自然な軽減が可能になる．以上から，直接推定法で求めた $r(a_k)$ の推定量に，特徴重み付け法を組み合わせることを提案する．実験では，言語資源から尤度比に基づき特定文脈を予測するタスクを設計し，重み付けが適切に作用する条件下で提案手法が有効なことを示す．

2 関 連 研 究

確率推定では，低・ゼロ頻度の離散要素に対する有効な推定法が多く提案されてきた．それらはスムージング法 [8] と呼ばれ，観測された要素の確率推定値から一定量を割り引き，それを未観測の要素に対する推定値へ分配する枠組みである．4 節で述べる提案手法では，その一つであるラプラススムージング [9] と同等の方法を頻度の補正に利用する．スムージング法は尤度比の推定にそのまま適用できないが，確率分布をスムージング法で推定して比を取ることはできる．しかし，確率分布の推定を工夫してその比を取るアプローチは，尤度比の推定誤差を大きくすることが報告されている [5]．それゆえ，尤度比の推定法は新たに考案する必要がある．

尤度比は確率分布の比で構成されるため，確率よりもその推定が困難である．尤度比の単純な推定法は，個々の確率分布を推定して比を取ること [10] だが，これは大きな推定誤差を生むことが明らかになっている．そのため，確率分布の推定を介さずに，尤度比を直接推定する手法が多く提案されてきた [6], [11], [12], [13]．これらの手法は連続空間での尤度比を推定の対象とし，標本空間から得られる要素も実数値を想定している．一方で，単語や文字等の離散要素から尤度比を直接推定する手法は提案されてこなかった．そこで近年，直接推定法を離散的な尤度比の推定に応用した手法 [5] が提案された．筆者らの知る限り，この手法は低頻度の離散要素から計算される尤度比にも，安定な推定値を与えられる唯一の方法である．しかし，ゼロ頻度の要素からは有効な推定値を算出できない．そこで本稿では，この手法をベースとして，ゼロ頻度の要素に対しても有効な尤度比の推定法を提案する．

ナイーブベイズ分類器に対しては，特徴間に仮定された条件付き独立性の軽減法が多く提案されてきた．例えば，1 節で述べた特徴重み付けの他に，構造拡張 [14]，特徴選択 [15]，インスタンス選択 [16]，インスタンス重み付け [17]，ファインチューニング [18] などの様々なアプローチが挙げられる．このうち，構造拡張は文献 [5] の手法と組み合わせることへの困難性があり，特徴選択とインスタンス選択は学習で使用できる観測頻度を減らしてしまう．そこで我々はこれらの問題を回避でき，かつ計算の効率性や応用の柔軟性にも優れている特徴重み付け法に焦点を当てた．特徴重み付け法は，重みを学習する方法の違いからフィルタ [7] とラッパー [19] に大別される．前者は一度の学習のみで重みを計算するのに対し，後者は分類性能のフィードバックを介して重みを最適化する．本稿では，重み付けする尤度比の推定量が調節を要する正則化パラメータを持つため，計算量が軽量なフィルタ方式を利用する．

3 尤度比推定法と特徴重み付け法

4 節で詳説する提案手法の要素技術となる，尤度比の直接推定法 [5]，ならびに特徴重み付け法 [7] を説明する．

3.1 尤度比の直接推定法

データの定義域を $D \subset \mathbb{U}^v$ とする．$\mathbb{U}^v$ は離散 v 次元空間である．いま，確率密度 $p_{\mathrm{de}}(x)$ を持つ確率分布に従う i.i.d. 標本と，確率密度 $p_{\mathrm{nu}}(x)$ を持つ確率分布に従う i.i.d. 標本

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

を得たとしよう．ここで，標本を構成する要素 x は単語や文字といった離散要素であり，要素の種類ごとに独立の空間が定義されることに注意する．すなわち，空間の次元数 v は存在しうる要素の種類数を意味する．これまでの先行研究に倣って，密度 $p_{\mathrm{de}}(x)$ が次の条件を満たすと仮定する．

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

これにより，すべての 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．

二乗損失に基づく直接推定法 unconstrained Least-Squares Importance Fitting（uLSIF）[6] では，$r(x)$ を次の線形和でモデル化する．

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

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_b)^{\mathrm{T}}$ 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．なお，b と $\{\varphi_l\}_{l=1}^{b}$ 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．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が，この基底関数では標本空間が離散であることを考慮できない．そこで，菊地ら [5] は要素の種類ごとに異なる，次の基底関数 $\{\delta_l\}_{l=1}^{v}$ を代用することを提案した．

$$\delta_l(x) = \begin{cases} 1 & x = x_{(l)} \\ 0 & x \neq x_{(l)} \end{cases} \tag{1}$$

インデックス 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，$x_{(l)}$ は v 種類存在する要素のうち，l 種類目の要素となる．式 (1) の基底関数を使用すると，$x_{(m)}$ $(1 \leqq m \leqq v)$ に対する推定モデルは次式となる．

$$\widehat{r}(x_{(m)}) = \sum_{l=1}^{v} \beta_l \delta_l(x_{(m)}) = \beta_m \tag{2}$$

uLISF では，推定モデル $\widehat{r}(x_{(m)})$ と真の尤度比 $r(x_{(m)})$ の二乗損失を最小化するパラメータ $\boldsymbol{\beta}$ を求める．その最適化問題は次式で与えられる [2]．

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^v} \left[\frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \widehat{\boldsymbol{H}} \boldsymbol{\beta} - \widehat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{\beta} + \frac{\lambda}{2} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta} \right] \tag{3}$$

$\mathbb{R}^v$ 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 $\frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$ を導入する．ここで λ $(\geqq 0)$ 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．$\widehat{\boldsymbol{H}}$ 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．

$$\begin{aligned} \widehat{H}_{l,l'} &= \frac{1}{n_{\mathrm{de}}} \sum_{i=1}^{n_{\mathrm{de}}} \delta_l(x_i^{\mathrm{de}}) \delta_{l'}(x_i^{\mathrm{de}}) \\ &= \begin{cases} \frac{1}{n_{\mathrm{de}}} f_{\mathrm{de}}(x_{(l)}) & (l = l') \\ 0 & (l \neq l') \end{cases} \end{aligned} \tag{4}$$

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．

$$\widehat{h}_l = \frac{1}{n_{\mathrm{nu}}} \sum_{j=1}^{n_{\mathrm{nu}}} \delta_l(x_j^{\mathrm{nu}}) = \frac{1}{n_{\mathrm{nu}}} f_{\mathrm{nu}}(x_{(l)}) \tag{5}$$

$f_*(x_{(l)})$ は，密度 $p_*(x)$ 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 (3) は拘束無し二次計画問題であり，その解は次式で解析的に求められる．

$$\widetilde{\boldsymbol{\beta}}(\lambda) = (\widehat{\boldsymbol{H}} + \lambda \mathbf{1}_v)^{-1} \widehat{\boldsymbol{h}}$$

$\mathbf{1}_v$ は要素が全て 1 の v 次元ベクトルである．そして式 (2)，(4)，(5) より，$\widehat{r}(x_{(m)})$ は次式で簡単に計算できる．

$$\begin{aligned} \widehat{r}(x_{(m)}) &= \widetilde{\beta}_m(\lambda) \\ &= (\widehat{H}_{m,m} + \lambda)^{-1} \widehat{h}_m \\ &= \left(\frac{1}{n_{\mathrm{de}}} f_{\mathrm{de}}(x_{(m)}) + \lambda \right)^{-1} \times \frac{1}{n_{\mathrm{nu}}} f_{\mathrm{nu}}(x_{(m)}) \end{aligned} \tag{6}$$

本来の uLSIF では，式 (3) の解が負の値を取ることがあり，尤度比の非負性を考慮して負の値をゼロに丸める必要があった．しかし上式は常に非負であるため，この式がそのまま $r(x_{(m)})$ の推定量になる．

式 (6) では，正則化パラメータ λ が $x_{(m)}$ の頻度に応じて推定値を低めに見積もる．λ がゼロのとき，この式は $r(x_{(m)})$ の分母・分子に相当する確率分布をそれぞれ最尤推定 [1] し，その比を取った結果に等しい．また，この式は尤度比の分母に相当する確率分布のみを補正する特殊な形式になっている．

3.2 ナイーブベイズ分類器に対する特徴重み付け法

ナイーブベイズ分類器は，強い独立性仮定とベイズの定理に基づく単純な確率的分類器である．いま，$A_1, A_2, \ldots, A_{\mathrm{N}}$ を N 個の特徴変数，インスタンス x を特徴ベクトル $\langle a_1, a_2, \ldots, a_{\mathrm{N}} \rangle$ とする．a_k $(1 \leqq k \leqq \mathrm{N})$ は A_k の取る値である．C をクラス変数，c を C の取る値とする．x を適切なクラスへと分類する問題を解くナイーブベイズ分類器は，次式で定義される．

$$\widehat{c}(x) = \arg\max_{c \in C} p(c) \prod_{k=1}^{\mathrm{N}} p(a_k \mid c) \tag{7}$$

$\widehat{c}(x)$ は x が分類されるクラスラベルである．条件付き確率 $p(a_k \mid c)$ は，各特徴 A_k がクラス c のもとで他の特徴と独立という条件付き独立性の仮定に基づく．この仮定は強力ながら実際には成り立たないことが多く，ナイーブベイズ分類器の性能を低下させる原因の一つとして知られている．

それゆえ，条件付き独立性の仮定を軽減する手法が多く提案されてきた．ここでは，特徴に応じて重み付けた条件付き確率を使用する，Feature Weighted Naive Bayes（FWNB）に着目する．FWNB は次式で定義される．

$$\widehat{c}(x) = \arg\max_{c \in C} p(c) \prod_{k=1}^{\mathrm{N}} p(a_k \mid c)^{W_k} \tag{8}$$

$W_k \in R^+$ は特徴 A_k に対する重みであり，独立性仮定の軽減に有効な W_k を計算する特徴重み付け法が多く提案されてきた．式 (8) と式 (7) との差異は重みの有無のみであり，同じ条件付き確率を式中で用いる．よって，a_k の頻度をそのまま利用できるため，分類器は訓練データの疎性が高い状況でも有効に作用すると考えられる．

本節では，特徴重み付け法の一つである Correlation-based Feature Weighting (CFW) filter [7] を紹介する．この手法は，分類に強く寄与する特徴がクラスと高い相関を持ち，他の特徴とは相関を持たないという前提に基づく．この前提に従い，特徴–クラス間の相関，および特徴間の平均した相互相関を求め，その差をシグモイド変換したものを特徴の重みとする．重みの学習アルゴリズムは単純かつ効率的で，計算された重みはナ

2：式 (3) の導出過程は uLSIF の原論文 [6] を参照のこと．

イーブベイズ分類器の性能向上に効果的なことが示されている．

CFW では相関を測る尺度として相互情報量を用いる．A_k $(1 \leqq k \leqq \mathrm{N})$ に対する特徴–クラス間の相互情報量 $I(A_k;C)$ は次式で計算される．

$$I(A_k;C) = \sum_{a_k}\sum_{c} p(a_k,c)\log\frac{p(a_k,c)}{p(a_k)p(c)}$$

同様に，A_k と $A_{k'}$ $(k \neq k')$ に対する相互情報量 $I(A_k;A_{k'})$ は次式で計算される．

$$I(A_k;A_{k'}) = \sum_{a_k}\sum_{a_{k'}} p(a_k,a_{k'})\log\frac{p(a_k,a_{k'})}{p(a_k)p(a_{k'})}$$

それぞれの確率は相対頻度を使用して計算される．$I(A_k;C)$ と $I(A_k;A_{k'})$ は下式でそれぞれ正規化される．

$$NI(A_k;C) = \frac{I(A_k;C)}{\frac{1}{\mathrm{N}}\sum_{k=1}^{\mathrm{N}} I(A_k;C)}$$

$$NI(A_k;A_{k'}) = \frac{I(A_k;A_{k'})}{\frac{1}{\mathrm{N}(\mathrm{N}-1)}\sum_{k=1}^{\mathrm{N}}\sum_{k'=1\wedge k'\neq k}^{\mathrm{N}} I(A_k;A_{k'})}$$

分類に強く寄与する特徴はクラスと高い相関を持ち，他の特徴とは相関を持たないという前提に基づき，A_k に対する重み D_k は次式で計算される．

$$D_k = NI(A_k;C) - \frac{1}{\mathrm{N}-1}\sum_{k'=1}^{\mathrm{N}} NI(A_k;A_{k'})$$

開区間 $(0,1)$ で値を取るよう，D_k は次式でシグモイド変換され，最終的な重み W_k が計算される．

$$W_k = \frac{1}{1+\mathrm{e}^{-D_k}}$$

重み W_k がゼロに近づくと式 (8) の条件付き確率 $p(a_k \mid c)$ は 1 に近づく．このとき，$p(a_k \mid c)$ が分類に与える影響は小さくなる．それに対して，W_k が 1 に近づくと $p(a_k \mid c)$ は本来の推定値に近づき，W_k の影響が小さくなる．

以上から CFW は，独立性仮定による悪影響を抑える重み付けになる．また，特徴 A_k とクラス変数 C が与えられれば，重み付けの対象が条件付き確率でなくとも使用できる．よって 4 節では，CFW を尤度比の推定に応用する．

4 提 案 手 法

特徴ベクトル $x = \langle a_1, a_2, \ldots, a_{\mathrm{N}}\rangle$ に対する尤度比

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

を推定する問題を考える．a_k $(1 \leqq k \leqq \mathrm{N})$ はベクトルの k 番目に位置する単語や文字などの離散値とし，その連なりである x は N-gram と呼ばれる．このとき，尤度比の簡易な推定法は x の頻度 $f_*(x)$ を使用して式 (6) の推定量 $\widehat{r}(x)$ を求めることである．しかし言語資源の頻度分布はべき乗則に従い，N が大きくなると $f_{\mathrm{nu}}(x) = 0$ となる x が増えるため，式 (6) の推定量が有効に作用しないことも多くなる．

前述の問題を緩和する一つの方法は，各特徴 A_k が他の特徴と独立であると仮定し，$r(x)$ を a_k に対する尤度比の積で表すことである．すなわち，この仮定の下で $r(x)$ は

$$r(x) = \prod_{k=1}^{\mathrm{N}} r(a_k)$$

と表される．しかしナイーブベイズ分類器の場合と同様，ここで仮定した独立性も実際には成立しないことが多い．そのため，この独立性仮定を軽減する手段が必要になる．

そこで，次式のように $r(a_k)$ を W_k で重み付けする．W_k の計算には 3.2 節で紹介した CFW を使用する．

$$r(x) \approx \prod_{k=1}^{\mathrm{N}} \widehat{r}(a_k)^{W_k} \tag{9}$$

$$\widehat{r}(a_k) = \left(\frac{1}{n_{\mathrm{de}}} f_{\mathrm{de}}(a_k) + \lambda\right)^{-1} \times \frac{1}{n_{\mathrm{nu}}} f_{\mathrm{nu}}(a_k)$$

加えて，$r(a_k)$ の推定に式 (6) の推定量 $\widehat{r}(a_k)$ を用いると，低頻度の a_k に対しても安定した推定値を算出できる．なお，ここでのクラス変数は $C = \{c_{\mathrm{de}}, c_{\mathrm{nu}}\}$ であり，それぞれの確率分布から得た標本を次のようにラベル付けする．

$$\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}} = \{x \mid c(x) = c_{\mathrm{de}}\}, \quad \{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}} = \{x \mid c(x) = c_{\mathrm{nu}}\}$$

$c(x)$ は x が属する真のクラスラベルである．これらの標本に基づいて W_k を計算する．

また，$f_{\mathrm{nu}}(a_k) = 0$ となる a_k が一つでも存在すると，$f_*(a_k)$ から計算される $\widehat{r}(a_k)$ がゼロとなり，その積である式 (9) もゼロになる．この問題を回避するため，a_k の頻度を補正して使用する．以上から，提案する推定量は次式となる．

$$r_{\mathrm{ours}}(x) = \prod_{k=1}^{\mathrm{N}} \widetilde{r}(a_k)^{W_k} \tag{10}$$

$$\widetilde{r}(a_k) = \left\{\frac{1}{n_{\mathrm{de}}+2}(f_{\mathrm{de}}(a_k)+1) + \lambda\right\}^{-1} \times \frac{1}{n_{\mathrm{nu}}+2}(f_{\mathrm{nu}}(a_k)+1)$$

$\frac{1}{n_*+2}(f_*(a_k)+1)$ は，密度 $p_*(x)$ を持つ確率分布をラプラススムージング [9] で推定した結果と等しい．他の補正法として n_* に a_k の種類数を足すことが考えられるが，疎性の高いデータでは a_k の種類数が n_* とほぼ等しくなり，補正前後の値が大きく異なる．そのため，提案手法では n_* に 2 を足すことにした．また，尤度比は確率分布の比であるため，頻度に加える補正が小さくても推定値が大きく変動すると考えられる．しかし，提案手法では正則化パラメータ λ $(\geqq 0)$ の導入により，推定値の大きな変動を抑制でき，補正が有効に機能する．提案手法ではそれぞれの $\widetilde{r}(a_k)$ が正則化パラメータを持つ．各パラメータは異なるもので，本来ならば A_k に応じて異なる最適値を設定することが望ましい．しかし，各パラメータを別々に変化させ，最適な組み合わせを探索することは計算コストが大きい．加えて n_{de} は A_k に依らず共通のため，本稿ではすべての λ が同じ値を持つと考え調節すべきパラメータを一つとした．

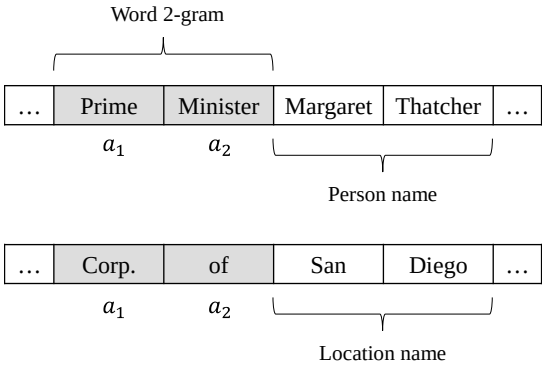


図 1: 固有表現とその左文脈の例（N = 2）.

5 評 価 実 験

低・ゼロ頻度の N-gram に対する提案手法の有効性を検証する．実験では，固有表現タグを付与したコーパスを使用して固有表現の左文脈を予測する．固有表現は地名と人名を採用し，文脈は固有表現の左に出現する単語 N-gram とする．固有表現およびその左文脈の例を図 1 に示す．人名の文脈では，人名の直前に “Mr.” や “Minister” といった敬称や役職を表す単語が出現しやすい．よって，人名の直前に位置する特徴 A_{N} が文脈の予測に大きな役割を果たす．一方で地名の文脈では，地名の直前に前置詞が出現しやすいが，これらは他の文脈でも出現しやすく，文脈の予測には他の単語との依存性も考慮する必要がある．よって人名と異なり，地名のケースでは文脈の予測に寄与する特徴の発見が難しい．性質が異なる二つの文脈を予測することは，提案手法のふるまいを解明する手助けになると考える．また，この実験では正解が一意に定義でき，手法を定量評価できる．さらに，言語資源の頻度分布はべき乗則に従うゆえ，N-gram の次数 N が大きくなると頻度の疎性が非常に高くなる．そのため，この実験は低・ゼロ頻度を多く扱うことになり，提案手法の有効性を検証するのに適している．

5.1 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 年版から次の手順で作成したデータを使用する．まず，Stanford Named Entity Recognizer（NER）[20] を用いてコーパスに固有表現タグを付与した．次に，タグ付けしたコーパスから記事をランダムサンプリングし，訓練データ，バリデーションデータ，テストデータへと割り振った．バリデーションデータとテストデータのサイズは常に 1,000 記事に固定し，訓練データのサイズは次の条件に応じて変化させた [4]．

- N-gram の次数 N : 2，4，6，8，10
- 訓練データのサイズ : 2,500，5,000，7,500，10,000 記事
- 固有表現 : 地名，人名

上の 2 条件は訓練データの疎性を操作し，最後の条件は予測する文脈を切り替えるために用いる．各条件からそれぞれ一つを選択し，その条件下で文脈の予測を試みる．紙面の都合により，N が 10，訓練データが 10,000 記事の場合に絞ってデータの性質を説明する．使用するデータの詳細を表 2 に示す．また，バリデーションデータ，テストデータが含む 10-gram に対する訓練データでの頻度分布を表 3 に示す．表 2 から，いずれのデータでも 10-gram の種類数と総頻度が近いことがわかる．さらに表 3 から，バリデーションデータとテストデータに含まれる 10-gram のうち，99%以上が訓練データに出現せず，残りも大半が低頻度なことがわかる．以上から，実験で扱うデータの疎性が確認できる．

表 2: 使用するデータ．N は 10，訓練データは 10,000 記事.

データ	全体の 10-gram		地名の左文脈		人名の左文脈	
	種類数	総頻度	種類数	総頻度	種類数	総頻度
Train	3,906,050	3,922,930	62,228	62,532	66,667	66,766
Valid	392,746	393,445	5,950	5,957	7,348	7,350
Test	394,850	395,145	5,713	5,716	7,520	7,522

表 3: バリデーションデータ，テストデータが含む 10-gram に対する訓練データ（10,000 記事）での頻度分布．ゼロ頻度は 10-gram が訓練データに含まれないことを意味する．

Valid		Test	
頻度	種類数	頻度	種類数
0	390,115	0	393,250
1	1,665	1	1,112
2	249	2	146
3	148	3	59
4	227	4	50
$\geqq 5$	342	$\geqq 5$	233
合計	392,746	合計	394,850

5.2 実 験 手 順

訓練データを N-gram $\langle a_1, a_2, \ldots, a_{\mathrm{N}} \rangle$ 単位に分割し，a_k $(1 \leqq k \leqq \mathrm{N})$ に対する訓練データ全体，および固有表現の左での出現頻度を数え上げる．テストデータに含まれる全ての N-gram x に対し，先ほど数え上げた頻度を利用して次の尤度比 $r(x)$ を推定する．

$$r(x) = \frac{p_{\mathrm{ne}}(x)}{p_{\mathrm{tr}}(x)}$$

$p_{\mathrm{tr}}(x)$ は x が訓練データの任意位置に出現する確率密度，$p_{\mathrm{ne}}(x)$ は x が固有表現の左に出現する確率密度を意味する．また提案手法では，尤度比の推定に重み W_k の計算を要する．訓練データを用いて W_k はあらかじめ計算しておく．

推定値の降順に N-gram x を種類ごとに順位付けして正誤判定する．判定ではテストデータ中で x が一度でも固有表現の左に出現すれば正解，そうでなければ不正解とする．判定した上位 8,000 件を用いて，横軸に x の順位，縦軸に再現率を取るランク—再現率曲線を描く．この曲線では，グラフの原点と曲線上のある点を結んだ直線の傾きが，その点における適合率

3 : https://catalog.ldc.upenn.edu/LDC2000T43

4 : 訓練データのサイズを変えるごとに，全データをリサンプリングしている．

に比例する．さらに，N が 10 あるいは訓練データのサイズが 10,000 記事の実験では，上位 8,000 件での F1 尺度も計算する．適合率，再現率，F1 尺度は次式で定義される．

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

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

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ここで，R はテストデータにおいて固有表現の左に出現する N-gram の集合，すなわち正解集合を意味する．

5.3 比較手法

実験では，低・ゼロ頻度 N-gram に対する提案手法の有効性のみならず，提案手法で導入される正則化パラメータ λ，および重み W_k それぞれの有効性，さらにはそれらの組み合わせることの有効性も検証する．このことを踏まえて提案手法を含む，以下の 4 手法を比較対象とした．なお，4 節の式 (10) に示した提案手法と同様，他の比較手法に対してもゼロ頻度への対策として補正した頻度を使用する [5]．

手法 1：ベースライン　尤度比 $r(a_k)$ の推定法として，確率分布の推定量をそれぞれ求めてその比を取る，間接推定法を用いる．手法の推定式は次式となる．

$$r_{\mathrm{B}}(x) = \prod_{k=1}^{\mathrm{N}} \bar{r}(a_k)$$

$$\bar{r}(a_k) = \left\{ \frac{1}{n_{\mathrm{tr}}+2}(f_{\mathrm{tr}}(a_k)+1) \right\}^{-1} \times \frac{1}{n_{\mathrm{ne}}+2}(f_{\mathrm{ne}}(a_k)+1)$$

手法 2：重み付け　手法 1 に対し，CFW [7] による重み付けを応用する．手法の推定式は次式となる．重み W_k は訓練データから計算でき，正則化パラメータの有無に依らない．それゆえ提案手法と共通の重みを用いる．

$$r_{\mathrm{W}}(x) = \prod_{k=1}^{\mathrm{N}} \bar{r}(a_k)^{W_k}$$

手法 3：正則化　菊地ら [5] の手法を用いて，尤度比 $r(a_k)$ を直接推定する．正則化パラメータ λ は後述の方法で設定する．手法の推定式は次式となる．

$$r_{\mathrm{R}}(x) = \prod_{k=1}^{\mathrm{N}} \widetilde{r}(a_k)$$

$$\widetilde{r}(a_k) = \left\{ \frac{1}{n_{\mathrm{tr}}+2}(f_{\mathrm{tr}}(a_k)+1) + \lambda \right\}^{-1} \times \frac{1}{n_{\mathrm{ne}}+2}(f_{\mathrm{ne}}(a_k)+1)$$

手法 4：正則化＋重み付け（提案手法）　手法 3 に対し，CFW による重み付けを応用する．λ は後述の方法で設定する．推定式は式 (10) に示したとおりである．

5：頻度の補正については予備実験にて有効性を確認した．具体的には，本稿と同様の実験で提案手法と式 (9) との性能を比較し，提案手法の優位性を示した．特徴重み付け法の利用に焦点を当てるため，本稿では実験結果の詳細を省略する．

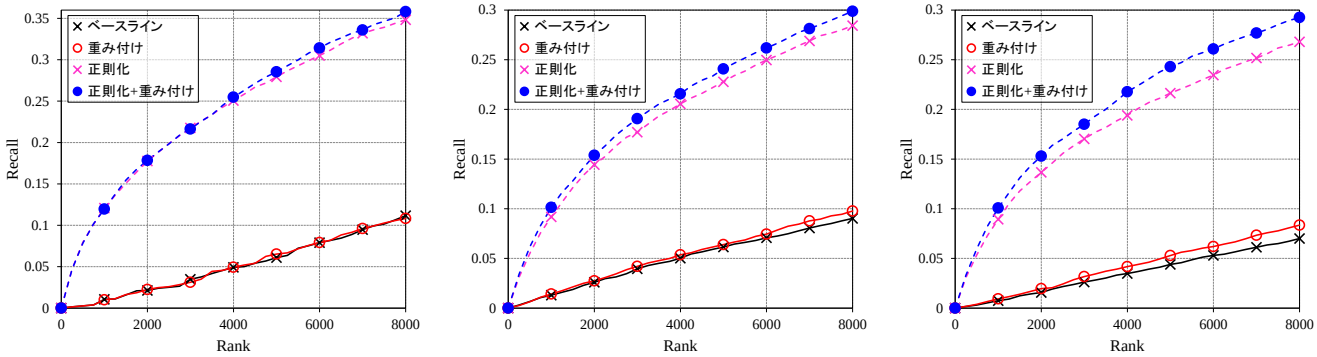
手法 3 と手法 4（提案手法）は，調節する正則化パラメータ λ を持つ．λ の最適値は次のように決定した．手法ごとに λ を $10^{-9}, 10^{-8}, \ldots, 10^{-1}$ と変化させ，それぞれについてバリデーションデータをテストデータとみなし，5.2 節の手順に従ってランク—再現率曲線を描いた．そして，その曲線下面積が最大となる λ の値を採用した．訓練データが 7,500 記事と 10,000 記事の場合は，重み W_k の有無と他の条件に依らず最適値が 10^{-5} となった．一方，訓練データが 2,500 記事と 5,000 記事の場合は最適値が 10^{-4} となった．よって，λ が訓練データのサイズ以外にはロバストという性質が見て取れた．

5.4 結果と考察

ランク—再現率曲線については，N が 2，6，10 かつ訓練データが 10,000 記事の場合に限り説明する．なお，他の実験条件でも，曲線から読み取れる手法間の優劣が同様の傾向を示すことを確認した．固有表現が地名および人名の曲線を図 2 と図 3 に示す．これらの曲線は横軸に N-gram の順位，縦軸に再現率を取る．また，グラフの原点と曲線上のある点を結んだ直線の傾きが，その点における適合率に比例する．ある順位で最高の再現率を持つ手法がその順位で最良の手法となる．ベースライン，重み付けのみの手法は，正則化を導入した二手法と比べて著しく性能が低い．これは低・ゼロ頻度の N-gram に対し，不当に高い推定値を付与したためと考える．重み付けのみの手法は，N が大きくなるとベースラインよりもやや高い再現率を示した．よって，正則化がない場合でも重み付けの有効性を確認した．一方，正則化を導入した二手法では大幅な性能向上が確認できた．これは低・ゼロ頻度の N-gram に対する尤度比の過大推定を防ぎ，安定した推定値を付与する正則化の作用によるものと考える．また，正則化と重み付けを組み合わせた提案手法は，N が大きくなるほど有効に作用した．特に，図 3(c) では予測対象が 10-gram でも，上位 8,000 件で 0.5 近くの高い再現率を示した．よって，正則化と特徴重み付けは相性が良く，尤度比の推定でそれらを組み合わせることの有効性が確認できた．

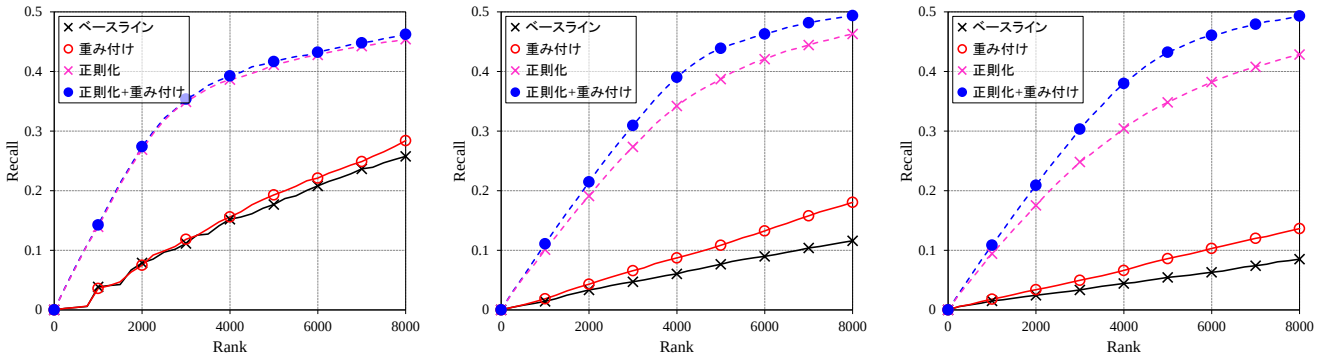
N を変化させた場合と，訓練データのサイズを変化させた場合の F1 尺度を図 4 と図 5 に示す．比較手法はランク—再現率曲線にて有効性を確認した二手法とした．図 4 から次のことがわかる．正則化のみの手法は N が 4 以上になると F1 値が単調減少する．それに対して，提案手法の F1 値は N の変化にロバストであり，N が大きくなるほど二手法の性能差が開いてゆく．図 4(b) に示す人名のケースでその傾向が顕著である．一方で図 5 からは次のことがわかる．N が 10 と大きい場合は，訓練データのサイズに依らず提案手法がほぼ一定の性能向上を示した．以上から，提案手法に導入された特徴重み付けは N が十分大きいと有効に作用し，その有効性は訓練データのサイズが小さくても保たれることが示唆された．

しかし，$\mathrm{N} = 2$ や固有表現が地名のケースでは，提案手法と正則化のみの手法との性能差が小さい．計算された重みに注目すると，固有表現が人名で N が 6 および 10 の場合は人名直前の特徴 A_6，A_{10} に大きな重み（それぞれ 0.820 と 0.890）が与えられた．一方で $\mathrm{N} = 2$ の場合は重み W_1 と W_2 が 0.414，



(a) N=2　(b) N=6　(c) N=10

図 2: 地名に対する文脈のランク—再現率曲線．訓練データは 10,000 記事である．



(a) N=2　(b) N=6　(c) N=10

図 3: 人名に対する文脈のランク—再現率曲線．訓練データは 10,000 記事である．

0.586 と共に小さく，人名直前という重要な特徴の特定が不十分であった．また，地名の場合は人名のように 0.8 を超える大きな重みがなく，重要な特徴の判別が困難であった．上記のケースでは重み付けが有効に作用せず，さらなる改善が必要になる．その方策として，特徴値 a_k とその出現位置の両方を考慮した重み付けを実現したい．a_k に対する重み付け法として，Term Weighting [21] 等の利用が考えられるが，これらを単純に適用するだけでは出現位置の情報を保持できない．したがって，両方の情報を保持した特徴重み付け法の開発，および尤度比推定への応用が今後の課題となる．

6 おわりに

低・ゼロ頻度 N-gram に対して有効な尤度比の推定法を提案した．提案手法は次の手順で定式化される．まず，N-gram を成す特徴 A_k $(1 \leqq k \leqq \mathrm{N})$ を個別に扱い，特徴値 a_k ごとに尤度比を推定する．次に，特徴間の独立性は実際に成立しないことが多いため，ナイーブベイズ分類器のための特徴重み付け法を，個々の尤度比に組み合わせる．これによって，特徴間に仮定された独立性による悪影響を軽減できる．最後に，重み付けした尤度比の積を取る．尤度比に基づく文脈予測の実験を行い，N が十分に大きい場合や予測に有効な特徴が特定しやすい場合において，提案手法の有効性を確認できた．今後の課題として，a_k に応じたより細かい重み付けの適用が考えられる．また，離散特徴を持つ分類問題への応用も期待できる．

文　献

[1] Scott Glover and Peter 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, 2015. IBISML2014-92.

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

[4] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[5] 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.

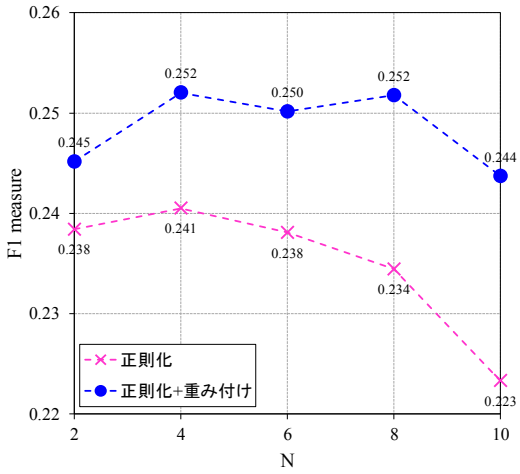
[6] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, Vol. 10, pp. 1391–1445, July 2009.

[7] Liangxiao Jiang, Lungan Zhang, Chaoqun Li, and Jia Wu. A correlation-based feature weighting filter for naive bayes. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31, No. 2, pp. 201–213, 2018.

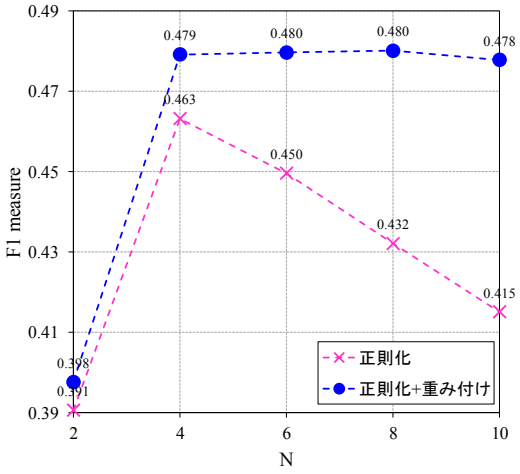
[8] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, Vol. 13, No. 4, pp. 359–394, 1999.

[9] George James Lidstone. 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.

[10] Wolfgang Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and semiparametric mod-*

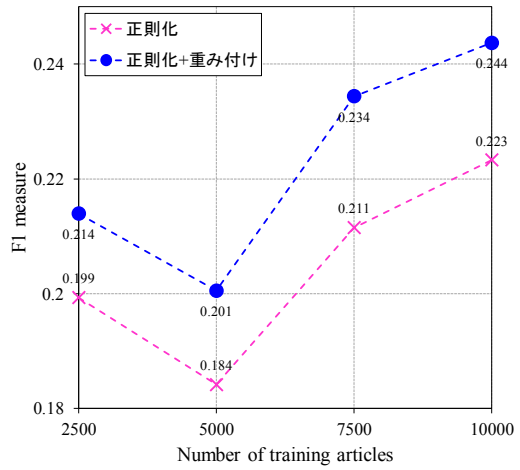


(a) 地名

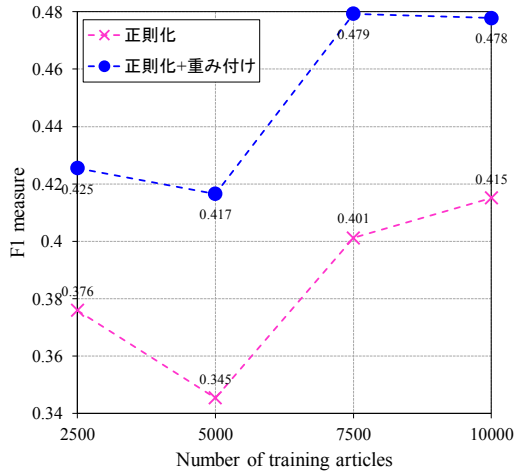


(b) 人名

図 4: 訓練データを 10,000 記事に固定し，N を変化させた F1 尺度.



(a) 地名



(b) 人名

図 5: N を 10 に固定し，訓練データのサイズを変化させた F1 尺度.

els. Springer Science & Business Media, 2012.

[11] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, pp. 601–608, 2007.

[12] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, pp. 81–88, 2007.

[13] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, pp. 1433–1440, 2008.

[14] Jia Wu, Shirui Pan, Xingquan Zhu, Peng Zhang, and Chengqi Zhang. SODE: Self-adaptive one-dependence estimators for classification. *Pattern Recognition*, Vol. 51, pp. 358–377, 2016.

[15] Bo Tang, Steven Kay, and Haibo He. Toward optimal feature selection in naive bayes for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 9, pp. 2508–2521, 2016.

[16] Eibe Frank, Mark Hall, and Bernhard Pfahringer. Locally weighted naive bayes. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI'03)*, pp. 249–256, 2003.

[17] Liangxiao Jiang, Dianhong Wang, and Zhihua Cai. Discriminatively weighted naive bayes and its application in text classification. *International Journal on Artificial Intelligence Tools*, Vol. 21, No. 1, 2012.

[18] Diab M. Diab and Khalil M. El Hindi. Using differential evolution for fine tuning naive bayesian classifiers and its application for text classification. *Applied Soft Computing*, Vol. 54, pp. 183–199, 2017.

[19] Nayyar A. Zaidi, Jesùs Cerquides, Mark J. Carman, and Geoffrey I. Webb. Alleviating naive bayes attribute independence assumption by attribute weighting. *Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 1947–1988, 2013.

[20] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 363–370, 2005.

[21] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620, 1975.