DEIM Forum 2022 K31-2(day3 p40)

正則化による尤度比推定法を応用した多値分類器の改良

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: †{kikuchi,ozono}@nitech.ac.jp

あらまし　不均衡データに対する分類問題を解くための一手法として，定義中に尤度比を含む Universal-set ナイーブベイズ分類器（UNB）が提案された．UNB において，尤度比推定は分類性能に寄与する重要な因子である．しかし，UNB で用いられる尤度比の推定法は，低頻度データの尤度比を過大に見積もってしまい，性能低下の原因となっている．我々は先行研究において，低頻度にも有効な尤度比推定の手法を提案した．この推定法では最適化における正則化を利用して，尤度比の過大推定を抑制できる．そこで本稿では，正則化による尤度比推定法を UNB に組み合わせることを提案する．不均衡データを用いた実験により，正則化パラメータがデータ量のバランスに応じて推定値を効果的に制御し，分類性能を向上させることを示す．

キーワード　尤度比推定，正則化，低頻度，Universal-set ナイーブベイズ分類器，不均衡データ

1　はじめに

データを適切なカテゴリへと振り分ける分類問題は，機械学習における代表的な課題の一つである．ナイーブベイズ分類器（NB）は古典的な確率的分類器で，良好な分類精度を保ち，分類効率が良く実装も容易といった利点から，昨今でもよく用いられている．一方で NB には様々な欠点もあり，これまでに多くの研究者が NB の改良に尽力してきた．NB が抱える問題の一つとして，不均衡データへの対処が挙げられる．実世界にあるデータでは，インスタンス（分類対象）が属するクラス（分類先）の占める割合が均衡ではないことがほとんどである．また，クラスの割合が極端に不均衡なデータもしばしば存在する．不均衡データを NB で扱うと，少数派クラスに多数のインスタンスが誤分類されることが知られている．加えて，少数派クラスに対してはモデルが疎になりやすい．

不均衡データを扱う方法として，補集合（クラスに属さないインスタンス）の利用がある．分類器で補集合を導入すると，モデルが疎になることを回避でき，分類精度の向上も期待できる．補集合を用いた分類器の一つとして，Universal-set ナイーブベイズ分類器（UNB）[1] が提案されており，

$$\widehat{c}(y) = \arg\max_{c \in C} \frac{p(c)}{p(\bar{c})} \prod_{k=1}^{n} r(w_k, c),$$

$$r(w_k, c) = \frac{p(w_k \mid c)}{p(w_k \mid \bar{c})}$$

と定義される．ここで $y = \langle w_1, w_2, \ldots, w_k, \ldots, w_n \rangle$ はインスタンスである．w_k は y の k 番目にある属性値で，数え上げが可能な文字や単語などの離散要素とする．また c はあるクラスを表し，$\bar{c}$ は c 以外の全クラスを表す．$\widehat{c}(y)$ は y が属すると分類器が予測したクラスである．UNB は不均衡データを扱うための分類器として提案されたが，我々は UNB を用いても不均衡データを精度良く分類できないと考えた．なぜなら，尤度比 $r(w_k, c)$ の推定法に問題があるからである．

表 1　尤度比の推定例．$\widehat{r}(x)$ の λ は 10^{-5} とした．

x	頻度				$r_{\mathrm{MLE}}(x)$	$\widehat{r}(x)$
	$\sum_x f_{\mathrm{de}}(x)$	$f_{\mathrm{de}}(x)$	$\sum_x f_{\mathrm{nu}}(x)$	$f_{\mathrm{nu}}(x)$		
x_{a}	10^7	2,000	10^4	100	50	47.6
x_{b}	10^7	20	10^4	1	50	8.3
x_{c}	10^7	20	10^4	2	100	16.7

UNB では w_k の相対頻度で各確率を求め，その比を取って尤度比を推定する．しかし我々は先行研究 [2] において，この一般的な推定法がまれな事象の尤度比を過大推定する問題を指摘した．具体例を用いて説明する．まず，推定する尤度比 $r(x)$ を

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

と定義する．続いて，上述の推定量 $r_{\mathrm{MLE}}(x)$ を

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

と定義する．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 $p_*(x)$ を持つ確率分布から得た x の頻度である．いま事象 x_{a} から x_{c} について，表 1 に示す頻度が与えられたとする．x_{a} と x_{b} に着目すると，頻度 $f_*(x_{\mathrm{a}})$ と $f_*(x_{\mathrm{b}})$ が大きく異なるのに対し，推定値は $\widehat{r}(x_{\mathrm{a}}) = \widehat{r}(x_{\mathrm{b}}) = 50$ と等しく大きな値である．しかし $f_{\mathrm{de}}(x_{\mathrm{b}}) = 1$ であり，この事象は偶然に起こったものかもしれない．少数派クラスには属性値 w_k がほとんど含まれず，含まれるとしてもその頻度は低い．ゆえに，低頻度事象の尤度比を過大推定することは，少数派クラスに対する分類スコアを不当に高く見積もることになり，多数のインスタンスをそのクラスへと誤分類する原因になる．また x_{b} と x_{c} に着目すると，$f_{\mathrm{nu}}(x_{\mathrm{b}})$ と $f_{\mathrm{nu}}(x_{\mathrm{c}})$ の差がわずか 1 であるのに対し，それぞれの推定値は 50，100 と大きく変動する．このことは，低頻度に対して UNB の分類スコアが不安定なこと意味し，この低頻度へのセンシティブさも UNB の分類精度を低下させる要因である．

我々は先行研究 [2] において，上述の問題を緩和する尤度比の推定量を提案した．この推定量は

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

と定義される．上式は二乗誤差を最小化する最適化の枠組みで導出される．λ（$\geqq 0$）は最適化の際に導入される正則化パラメータであり，λ が頻度に応じて推定値を低め（保守的）に見積もる．表 1 に示すように，高頻度から計算される $\widehat{r}(x_{\mathrm{a}})$ は 50 に近いのに対し，低頻度から計算される $\widehat{r}(x_{\mathrm{b}})$ は 50 よりもはるかに低い（それぞれ 47.6，8.3）．また，$\widehat{r}(x_{\mathrm{c}})$ も低頻度から計算されるため，16.7 と 100 よりもはるかに低い値となる．したがって $\widehat{r}(x)$ を用いると，尤度比の過大推定を抑え，低頻度にロバストな推定を行える．

以上を踏まえて本稿では，UNB に尤度比の保守的な推定量を組み合わせることを提案する．また，単に組み合わせるだけではなく，クラス毎に異なる正則化パラメータを用意し，データ中のクラスバランスに応じて個々のパラメータ値を変更する．実験では，コーパスから抽出した固有表現の出現文脈を多値分類することを試みる．そして実験結果から以下のことを明らかにする．UNB は訓練データに少数派クラスがある場合，多数のインスタンスをそのクラスへ誤分類してしまう．結果として，ときには古典的な NB をも大幅に下回る分類精度を示す．提案手法は少数派クラスに大きめな正則化パラメータを設定し，分類スコアの不当な高まりを抑える．結果として，訓練データに極端な少数派クラスがある場合でも良好な分類精度を保ち，UNB の欠点を緩和できる．

2 関 連 研 究

不均衡データを分類するための様々な手法が提案されている．それらの手法は，データレベルのアプローチとモデルレベルのアプローチに大別される．データレベルのアプローチとして，アップサンプリング [3], [4] とダウンサンプリング [5], [6] が有名である．これらのサンプリング法では，入力データ間の距離を測定し，その距離を利用して標本サイズを人為的に増減する．データレベルのアプローチは分類に用いるモデルの種類によらず利用できるが，本稿で扱うような離散値を含むデータでは，入力データ間の距離の定義が難しく利用範囲が限られる．モデルレベルのアプローチとして，補集合を用いた NB が提案された [1], [7][1]．提案手法は，補集合を利用したモデルレベルのアプローチに属する．また，Cost-Sensitive Learning [8] を利用した NB も提案された [9], [10], [11]．これらの NB では，各クラスの分類失敗に対して異なる “コスト” を導入し，分類時の期待コストが最小となるように訓練や分類を行う．適切なコストの設定には専門家の知識を要するが，提案手法にもコストを導入して正則化パラメータを調節できれば，各クラスに対する分類精度の細かい調整が容易になるかもしれない．

1：UNB 以外の NB との比較も行い，提案手法の有効性を確認した．比較結果は本稿の付録として掲載している．

確率分布をそれぞれ推定し，その比を取るという尤度比の推定法は，大きな推定誤差を生むことが明らかになっている [12]．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存在する．これまでに Kernel Mean Matching による手法 [13]，ロジスティック回帰による手法 [14]，カルバック・ライブラー情報量を用いた手法 [15]，最小二乗法による手法 [16] などが提案された．しかしこれらの手法は，連続的な標本空間で定義される尤度比を推定対象とし，扱う標本の要素も連続値を想定している．そこで我々は，最小二乗法による手法 unconstrained Least-Squares Importance Fitting（uLSIF）[16] で用いる基底関数を変更し，離散的な標本空間で定義される尤度比を推定できるようにした [2]．さらに我々は，単語 N-gram を構成する個々の単語に対して尤度比を推定し，それらの積を取ることで訓練データにない N-gram にも尤度比を推定できる手法を提案した [17]．そして，二値分類において有効性を示した．この推定法は NB から着想を得ており，尤度比推定に先行研究 [2] の成果を利用している．本稿ではこの推定法の発想を多値分類にも応用する．

3 前 提 知 識

提案手法の導入に必要となる，確率的分類器と尤度比の保守的な推定法について述べる．

3.1 ナイーブベイズ分類器

文字や単語などの離散要素 w_k（$k = 1, 2, \ldots, n$）からなるインスタンスを $y = \langle w_1, w_2, \ldots, w_n \rangle$ とする．また C をクラス変数，$c \in C$ をクラスとする．このとき，y が属するとナイーブベイズ分類器（NB）が予測するクラス $\widehat{c}(y)$ は

$$\widehat{c}(y) = \arg\max_{c \in C} p(c \mid y)$$

と定義される．$p(c \mid y)$ に対してベイズの定理を適用すると，

$$\begin{aligned}\widehat{c}(y) &= \arg\max_{c \in C} \frac{p(y \mid c)p(c)}{p(y)} \\ &= \arg\max_{c \in C} p(y \mid c)p(c) \\ &= \arg\max_{c \in C} p(w_1, w_2, \ldots, w_n \mid c)p(c)\end{aligned}$$

が得られる．ここで，w_k がクラス c の下で他の離散要素と条件付き独立であることを仮定すると

$$p(w_1, w_2, \ldots, w_n \mid c) = \prod_{k=1}^{n} p(w_k \mid c)$$

が成り立つ．以上より，NB は

$$\widehat{c}(y) = \arg\max_{c \in C} p(c) \prod_{k=1}^{n} p(w_k \mid c) \tag{1}$$

と定式化される．NB は分類タスクでよく用いられるが，クラス間のサイズが大きく異なる場合，分類精度が著しく低下する問題点がある．そこで，補集合（クラスに属さないインスタンス）を用いた分類器が提案された．

3.2 Universal-set ナイーブベイズ分類器

補集合を利用した確率的分類器である，Universal-set ナイーブベイズ分類器（UNB）[1] について述べる．$p(c \mid y)$ と $p(\bar{c} \mid y)$ について

$$p(c \mid y) + p(\bar{c} \mid y) = 1$$

が成り立つ．ただし，クラス $\bar{c}$ は c を除いた全クラスを意味する．$p(c \mid y)$ と $p(\bar{c} \mid y)$ に対して，それぞれベイズの定理を適用すると上式は

$$\frac{p(y \mid c)p(c)}{p(y)} + \frac{p(y \mid \bar{c})p(\bar{c})}{p(y)} = 1$$

と変形できる．この式を $p(y)$ について解くと

$$p(y) = p(y \mid c)p(c) + p(y \mid \bar{c})p(\bar{c})$$

が得られ，これを $p(c \mid y)$ に代入すると

$$\begin{aligned} p(c \mid y) &= \frac{p(y \mid c)p(c)}{p(y)} \\ &= \frac{p(y \mid c)p(c)}{p(y \mid c)p(c) + p(y \mid \bar{c})p(\bar{c})} \\ &= \frac{1}{1 + \frac{1}{\alpha}} \end{aligned}$$

が得られる．ただし α は

$$\alpha = \frac{p(y \mid c)p(c)}{p(y \mid \bar{c})p(\bar{c})} = \frac{p(w_1, w_2, \ldots, w_n \mid c)p(c)}{p(w_1, w_2, \ldots, w_n \mid \bar{c})p(\bar{c})}$$

である．ここで $p(c \mid y)$ が最大となるのは，α が最大となるときである．また w_k に対して，クラス c および $\bar{c}$ の下での条件付き独立性を仮定する．以上より，UNB は

$$\widehat{c}(y) = \arg\max_{c \in C} \frac{p(c)}{p(\bar{c})} \prod_{k=1}^{n} r_{\mathrm{MLE}}(w_k, c) \tag{2}$$

と定式化される．ただし上式では，尤度比 $r(w_k, c)$ を

$$r_{\mathrm{MLE}}(w_k, c) = \frac{\widehat{p}(w_k \mid c)}{\widehat{p}(w_k \mid \bar{c})} \tag{3}$$

として推定する．クラス不均衡性による分類性能の悪化を軽減する目的で，UNB では補集合を使用する．また UNB では，式 (3) に示すように w_k の相対頻度によって確率推定量 $\widehat{p}(w_k \mid c)$ と $\widehat{p}(w_k \mid \bar{c})$ を計算し [2]，その比を取って推定量 $r_{\mathrm{MLE}}(w_k, c)$ を得る．この推定法は一般的かつ単純だが，低頻度に基づく尤度比を過大に見積もることがよくある．クラス不均衡性のある分類問題では，クラス間で w_k の頻度が大きく異なり，少数派クラスにおける w_k の頻度は他クラスの頻度よりも低くなる．このとき，少数派クラスに対する尤度比が他クラスの尤度比と比較して不当に高くなり，多くのインスタンスが少数派クラスに誤分類されてしまう．ゆえに，尤度比の過大推定を抑制する何らかの工夫が必要となる．

2：ただし観測頻度をそのまま用いた場合，確率推定値がゼロになる w_k が一つでもあると，式 (2) における尤度比の積がゼロや無限大になる．これを防ぐため，実際には正の値を各頻度へ加算する補正を行う．

3.3 尤度比の保守的な推定量

尤度比推定の問題設定を説明する．あるデータが含む離散要素 x の集合を $D \subset \mathbb{U}$ とする．$\mathbb{U}$ 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呼ばれる．いま，確率密度 $p_{\mathrm{de}}(x)$ を持つ確率分布に従う標本，確率密度 $p_{\mathrm{nu}}(x)$ を持つ確率分布に従う標本

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

をそれぞれ得たとする．ここで言う離散要素 x とは，文字や単語等の言語要素を表す．先行研究にならい，

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

を満たすと仮定する．この仮定により，全ての x に対して尤度比を定義できる．本節では，尤度比

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

を確率分布推定を介さず，二つの標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$，$\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ を用いて直接推定する問題を扱う．

最小二乗法による直接推定法 unconstrained Least-Squares Importance Fitting（uLSIF）[16] は，推定モデル $\widehat{r}(x)$ を

$$\widehat{r}(x) = \sum_{l=1}^{b} \beta_l \varphi_l(x) \tag{4}$$

と定義する．ここで，$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_b)^T$ 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．オリジナルの uLSIF は，連続空間上で定義される尤度比を扱う．そのため，連続的な標本空間の構造を活用する目的で，ガウスカーネルによる基底を使用する．しかし本稿では，文字や単語といった離散要素を扱い，尤度比も離散空間上に定義される．それゆえ，ガウスカーネルが有用ではない．そこで，我々が先行研究 [2] においてが定義した基底 $\{\delta_l\}_{l=1}^{v}$

$$\delta_l(x) = \begin{cases} 1 & \text{if } x = x_{(l)} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

を代用する．$\delta_l(x)$ は離散要素の種類ごとに定義される．l は要素の種類を指定する添え字であり，$x_{(l)}$ は v 種類存在する要素のうち l 種類目の要素を表す．この基底を用いると，離散空間上の尤度比に対して単純な推定式を導出できる．ここで b は v に置換される．式 (5) を式 (4) に代入し，推定モデル $\widehat{r}(x)$ と真の尤度比 $r(x)$ との二乗誤差を最小化する $\boldsymbol{\beta}$ を求める [3]．$x_{(m)}$ $(m = 1, 2, \ldots, v)$ に対する最終的な推定量 $\widehat{r}(x_{(m)})$ は

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

となる．$f_*(x_{(m)})$ は，密度 $p_*(x_{(m)})$，$* \in \{\mathrm{de}, \mathrm{nu}\}$ を持つ確率分布からサンプリングされた $x_{(m)}$ の頻度であり，$n_* = \sum_x f_*(x)$ である．上式では，正則化パラメータ λ（$\geqq 0$）によって，頻度

3：$\boldsymbol{\beta}$ の導出過程は文献 [2] を参照のこと．

に応じて推定値を低め（保守的）に見積もる．ただし，尤度比の積がゼロや無限大になることを防ぐため，頻度に微小な補正をした推定量

$$\widetilde{r}(x_{(m)}) = \left\{ \frac{f_{\mathrm{de}}(x_{(m)}) + 1}{n_{\mathrm{de}} + 2} + \lambda \right\}^{-1} \frac{f_{\mathrm{nu}}(x_{(m)}) + 1}{n_{\mathrm{nu}} + 2} \qquad (6)$$

を提案手法の尤度比推定に利用する．なお，補正した確率推定量 $\frac{f_*(x_{(m)})+1}{n_*+2}$ は，密度 $p_*(x_{(m)})$ を持つ確率分布を $x_{(m)}$ が出現するか否かのベルヌーイ試行による確率分布と捉え，事前分布を一様分布としたときの事後期待値と等しい．

4 提案手法

UNB では，相対頻度により各確率を推定し，その比を取って式 (3) の推定量 $r_{\mathrm{MLE}}(w_k, c)$ を得る．しかし $r_{\mathrm{MLE}}(w_k, c)$ は，低頻度に基づく尤度比を過大に見積もる．3.2 節でも述べたように，$r_{\mathrm{MLE}}(w_k, c)$ を用いると多くのインスタンスが少数派クラスに誤分類されてしまう．これを防ぐため我々は，UNB の尤度比推定に式 (6) を利用する．よって提案手法は

$$\widehat{c}(y) = \arg\max_{c \in C} \frac{p(c)}{p(\bar{c})} \prod_{k=1}^{n} \widetilde{r}(w_k, \lambda_c) \qquad (7)$$

と定式化される．提案手法では，クラスごとに異なる複数の正則化パラメータ $\boldsymbol{\lambda} = \{\lambda_c \mid c \in C\}$ を用いる．これにより，クラスバランスに応じて推定値の大きさを制御し，多数のインスタンスが少数派クラスに誤分類されることを防ぐ．そのために，推定量 $\widetilde{r}(w_k, \lambda_c)$ にはクラス c 自体ではなく，クラスごとの正則化パラメータ λ_c を引数として与えている．

正則化パラメータ $\boldsymbol{\lambda}$ の決定方法を述べる．ここでは，$\lambda_c \in \boldsymbol{\lambda}$ が取る値の候補を $\Theta = \{10^{-9}, 10^{-8}, \ldots, 10^{-1}\}$ とする．このとき，分類結果に基づくある評価関数 J を最大化する $\lambda_c^{'}$ の組み合わせ $\boldsymbol{\lambda}^{'} = \{\lambda_c^{'} \mid c \in C\}$ を求める．M クラス分類問題を想定すると，この最適化問題は

$$\boldsymbol{\lambda}^{'} = \arg\max_{\boldsymbol{\lambda} \in \Theta^{\mathrm{M}}} J(\mathrm{MLC},\ \boldsymbol{\lambda})$$

として定式化される．MLC は $\boldsymbol{\lambda}$ をパラメータとして持つ何らかの多値分類器であり，ここでは式 (7) に示す我々の分類器とする．本稿では不均衡データを扱うことを踏まえ，評価関数 J をデータの分類結果から計算される Macro-F1 とした．少数派クラスに多数のインスタンスが誤分類される場合，Macro-F1 は低くなることが想定される．したがって，Macro-F1 を最大化する正則化パラメータを用いると，提案手法は各クラスの分類性能が平均的に良好な分類器になると考える．

ただし，正則化パラメータの取りうる値の組み合わせ総数は 9^{M} 通りであり，クラス数 M が大きいときは組み合わせの全探索が困難なことに注意する．そこでパラメータの最適値を近似的に求める一例として，最適化手法の一種である差分進化法 [18] を利用する [4]．差分進化法の実行に必要な各種パラメータは次のように設定した．

- 最大進化世代 MaxGen = 50
- 解の候補数（個体群サイズ）NP = 30
- 変動係数 F = 0.8
- 交叉確率 CR = 0.6

最適化の目的関数である適応度関数を Macro-F1 とし，関数の値を最大化するパラメータ $\boldsymbol{\lambda}^{'}$ を最適値とみなす．なお，Macro-F1 を算出するために開発データを用意し，評価データとみなして分類問題を解いた．実験に用いるデータは 5.1 節で説明する．

5 評価実験

固有表現（コーパスに出現する地名や人名などの固有名詞）の出現文脈を 6 あるいは 7 クラス分類する実験を行い，提案手法の有効性を明らかにする．対象とする固有表現は LOCATION, ORGANIZATION, DATE, MONEY, PERSON, PERCENT, TIME の 7 種類である．また出現文脈（インスタンス）は，固有表現の左に位置する単語 10-gram とする．本実験を行う理由は次の三点である．第一に，言語要素は種類が豊富な反面，低頻度のものが多い．この性質により，尤度比の過大推定を抑制する正則化パラメータの効果を検証できる．第二に固有表現は出現のしやすさが大きく異なる．5.1 節で後述するように，ORGANIZATION の文脈が 91,892 回も出現するのに対し，TIME の文脈は 840 回しか出現しない．このように，非常に不均衡なデータを用いることからも，本実験は提案手法の有効性検証に適する．第三に固有表現の出現文脈は一意に定まり，分類器の定量評価が可能となる．

5.1 実験データと実験条件

実験データはウォール・ストリート・ジャーナルコーパス [5] の 1987 年版をもとに作成した．まず，コーパスからランダムに 12,000 記事を抽出し，10,000 記事，1,000 記事，1,000 記事を訓練，開発，評価用にそれぞれ分配した．次に，Stanford Named Entity Recognizer（Stanford NER）[6] [19] を用いて記事に固有表現タグを付与し，出現文脈（インスタンス）を抽出した．訓練用の記事が含むインスタンスを図 1 に示す．この図から分かるように，TIME の文脈は他の文脈と比較して極端に少ない．開発，評価用の 1,000 記事から抽出したインスタンス集合をそれぞれ開発データ，評価データとする．

次の訓練データを用意し，どの訓練データを用いるかを実験条件とする．

- クラス均衡，クラス不均衡な人工データ
- 実データ

人工データを作成する際は，指定したクラスサイズになるよう訓練用の記事からインスタンスをランダムサンプリングした．なおクラスサイズとは，クラスに属するインスタンスの総数を意味する．クラス均衡な人工データとして，各クラスサイズを 15,000，あるいは 1,500 に揃えた二種類のデータを用意した．クラス不均衡な訓練データとして，あるクラスサイズを 1,500,

4：今回は差分進化法を用いるが，貪欲法などの他手法でも探索可能である．

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

6：https://nlp.stanford.edu/software/CRF-NER.html

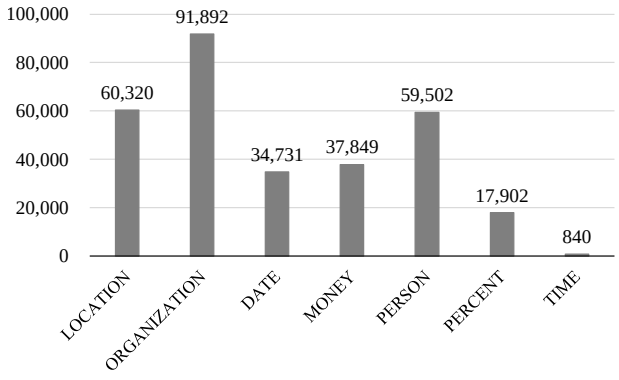

図 1　訓練用の 10,000 記事が含むインスタンス．

他のクラスサイズを 15,000 としたデータ，およびあるクラスサイズを 15,000，他のクラスサイズを 1,500 としたデータを用意した [7]．訓練用の記事が含む全インスタンスの集合を実データとした．実データはコーパス中の文脈に対する頻度の偏りを反映したデータである．すなわち，図 1 は実データにおける各クラスサイズであり，実データは非常に不均衡なデータである．

5.2 実験手順

使用する訓練データを事前に定め，以下の手順で実験を行う．訓練データから分類スコアの計算に必要な頻度を数え上げる．提案手法では，開発データを用いてクラスごとの正則化パラメータ λ_c を決定する．評価データにある全種類のインスタンスに対して分類を行う．分類器に対する分類性能を算出する．分類性能の指標として再現率，適合率，F1 値の Macro 平均，正解率の Micro 平均を採用する．

人工データによる実験では，上記の手順を 6 回繰り返し，各指標の平均値を算出する．例えば，再現率の平均値は

$$R_{\mathrm{avg}} = \frac{\sum_{t=1}^{6} R_t}{6}$$

として計算する．R_t は t 回目の試行における再現率の Macro 平均である．同様に P_t，F_t を t 回目の試行における適合率，F1 値の Macro 平均，A_t を t 回目の試行における正解率の Micro 平均とする．そして P_{avg}，F_{avg}，A_{avg} を計算する．なお，試行ごとにインスタンスのランダムサンプリングからやり直し，異なった訓練データを用いて実験する．クラス不均衡な人工データによる実験では，サイズの異なる 1 クラスを LOCATION, ORGANIZATION, . . . , PERCENT と順番に変化させて 6 回試行する．

5.3 比較手法

比較手法として NB，UNB，提案手法を用いる．

手法 1：NB　UNB において補集合を用いることの効果を検証するため，比較手法として NB を用いる．NB は 3.1 節の式 (1) として定義される．なお，$p(w_k \mid c)$ の推定にはラプラススムージングを利用する．

7：図 1 に示すように，TIME の文脈は人工データを作成するには少なすぎる．そのため，人工データを用いた実験では，TIME を除く 6 クラス分類を行った．

表 2　クラス均衡な人工データによる分類結果．それぞれの評価値（再現率 R_{avg}，適合率 P_{avg}，F1 値 F_{avg}，正解率 A_{avg}）は 6 回試行した平均値である．

訓練データの各クラスサイズ	分類器	Macro			Micro
		R_{avg}	P_{avg}	F_{avg}	A_{avg}
15,000	NB	.603	.552	.560	.568
	UNB	.610	.559	.572	.577
	提案手法	.607	.599	.597	.600
1,500	NB	.522	.477	.480	.489
	UNB	.529	.481	.488	.497
	提案手法	.508	.578	.526	.533

表 3　クラス不均衡な人工データによる分類結果（あるクラスサイズ：1,500，他のクラスサイズ：15,000）．それぞれの評価値は 6 回試行した平均値である．

分類器	Macro			Micro
	R_{avg}	P_{avg}	F_{avg}	A_{avg}
NB	.538	.494	.498	.507
UNB	.410	.682	.397	.376
提案手法	.572	.586	.561	.555

表 4　クラス不均衡な人工データによる分類結果（あるクラスサイズ：15,000，他のクラスサイズ：1,500）．それぞれの評価値は 6 回試行した平均値である．

分類器	Macro			Micro
	R_{avg}	P_{avg}	F_{avg}	A_{avg}
NB	.427	.468	.399	.404
UNB	.508	.511	.457	.476
提案手法	.483	.577	.482	.501

手法 2：UNB　UNB は 3.2 節の式 (2) として定義される．ただし，w_k の相対頻度に対して分母に 2，分子に 1 を加算する補正を行う [8]．UNB における尤度比の推定量は，式 (6) の λ をゼロとした結果に等しい．

手法 3：提案手法　提案手法は 4 節の式 (7) として定義される．正則化パラメータ λ_c の探索に差分進化法を用い，クラスごとに異なる最適値 $\lambda_c^{'}$ を設定する．

5.4 人工データによる実験結果

人工データによる分類結果を表 2〜4 に示す．各表において評価値の最大値を下線で強調した．クラス均衡な場合（表 2）では，UNB が NB よりも優れており，補集合を用いることの有効性を確認した．正則化でクラス毎に推定値を調整できる提案手法は，再現率が他二手法と比較してわずかに劣るが，適合率が優れている．ゆえに提案手法は，F1 値や正解率の観点から最良の性能を示している．クラス不均衡な場合（表 3，表 4）では，クラス均衡な場合と比較して手法間に大きな性能差が見られた．表 3 から分かるように，一つだけ少数派クラスがあるとき，UNB は NB をも下回る低い性能を示している．これは少

8：確率推定量の補正にはスムージング法を用いることが一般的である．しかし，スムージング法による推定値を尤度比推定に使用してしまうと，補正前後の尤度比が大きく異なり過大推定の原因にもなる [2]．したがって本稿では，頻度に最低限の微小な補正のみを行う．

表 5　差分進化法で求めた正則化パラメータの最適値 λ_c'（左は全クラスを用いた実験，右は TIME を除いた実験）．

Class	λ_c'	Class	λ_c'
LOCATION	10^{-6}	LOCATION	10^{-9}
ORGANIZATION	10^{-9}	ORGANIZATION	10^{-9}
DATE	10^{-5}	DATE	10^{-5}
MONEY	10^{-5}	MONEY	10^{-5}
PERSON	10^{-5}	PERSON	10^{-5}
PERCENT	10^{-4}	PERCENT	10^{-5}
TIME	10^{-3}		

表 6　実データによる分類結果（全クラス）．下線で強調された数値は各指標の最大値である．

分類器	Macro			Micro
	再現率	適合率	F1 値	正解率
NB	.546	.510	.483	.529
UNB	.299	.713	.241	.160
提案手法	.522	.678	.540	.626

表 7　実データによる分類結果（TIME 除外）．下線で強調された数値は各指標の最大値である．

分類器	Macro			Micro
	再現率	適合率	F1 値	正解率
NB	.609	.574	.571	.582
UNB	.610	.578	.564	.568
提案手法	.624	.624	.610	.617

数派クラスの尤度比が過大推定され，多くのインスタンスがそのクラスへと誤分類されたためと考える．そのような状況でも提案手法は良好な性能を保っており，正則化パラメータの有効性が示唆された．表 4 から分かるように，一つだけ多数派クラスがあるとき，UNB は NB よりも良い性能を示した．提案手法は再現率が UNB よりも若干劣るものの，適合率が良好であり，結果として F1 値や正解率の観点から最良の性能を示した．

5.5　実データによる実験結果

5.1 節の図 1 で示した実データを訓練に用いた場合の分類結果について述べる．提案手法で用いた正則化パラメータの最適値 λ_c' を表 5 左に示す．この表と図 1 を比較すると，多数派クラスには小さい値が，少数派クラスには大きい値が付与されたことが分かる．よって，提案手法はクラスサイズに応じて正則化パラメータを調節する．表 6 に示す分類結果から分かるように，UNB は F1 値と正解率が低く，効果的な分類ができていない．一方で，提案手法は F1 値と正解率が最良であり，正則化パラメータの有効性が示唆された．UNB と提案手法の性能差を詳細に分析するため，クラスごとの再現率，適合率，F1 値を図 2 に示す．この図から分かるように，UNB は TIME の再現率が高く，適合率と F1 値がゼロに近い．また，TIME 以外のクラスでは適合率のみが高く，再現率と F1 値が低い．これは，UNB が多くのインスタンスを TIME へと誤分類したこと意味する．それに対し，提案手法は TIME の F1 値が低いが，他クラスの F1 値は向上し，少数派クラスへの誤分類を抑制できた．

TIME は極端な少数派クラスである．そこで訓練，開発，評価データから TIME を取り除き，各分類器で 6 クラス分類も行った．正則化パラメータの最適値 λ_c' を表 5 右に示す．TIME を含めた全クラスの場合よりも最適値のばらつきは小さいが，今回も多数派クラスには小さい値が，少数派クラスには大きい値が付与される傾向があった．表 7 に示す分類結果から分かるように，TIME を除いて分類すると UNB の性能は大きく向上する．しかし，UNB の F1 値と正解率は NB よりも低く，UNB は改良が必要なことが示唆された．一方で，提案手法は最良の性能を持ち，実データを用いた実験でも有効性を確認できた．UNB と提案手法におけるクラスごとの再現率，適合率，F1 値を図 3 に示す．この図から分かるように，二手法間には図 2 のような大きな性能差は見られない．提案手法について図 2(b) と図 3(b) を見比べる．PERCENT では，再現率と適合率に大きな変化があるが，F1 値はさほど変わらない．また他クラスではいずれの指標も大きな変化がない．この結果は，分類対象に極端な少数派クラスを含む場合でも，提案手法には多数派クラスの分類性能をあまり落とさない利点があることを示唆した．

6　おわりに

本稿では，我々の先行研究である尤度比の保守的な推定法を UNB へ組み合わせ，新たな分類器を提案した．提案手法は，尤度比推定で導入される正則化パラメータ λ_c を用い，クラスバランスに応じて分類のスコアを調節する．不均衡データによる多値分類の実験において，極端な少数派クラスがある場合，UNB は尤度比を過大推定してしまい，ベースラインの NB をも下回る分類性能を示した（Macro-F1：0.241）．一方で提案手法は，尤度比の過大推定を正則化パラメータで抑制し，最良の性能を達成した（Macro-F1：0.540）．

しかし図 2 に示すように，提案手法でも少数派クラスの F1 値は低い．実際の分類問題では，他クラスの分類性能を多少犠牲にしても，少数派クラスの性能を向上させたいことがある．提案手法はクラス毎に分類のスコアを調節できるため，λ_c の探索法を変更すれば，少数派クラスの分類性能を向上できるだろう．一例として Cost-Sensitive Learning を用い，少数派クラスの分類失敗に大きなコストを定義し，期待コストが最小となるように λ_c を学習することが考えられる．以上のような，実用に向けた提案手法の拡張が今後の課題である．

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] K. Komiya, Y. Ito, and Y. Kotani. New naive bayes methods using data from all classes. *International Journal of Advanced Intelligence*, Vol. 5, No. 1, pp. 1–12, 2013.

[2] 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.

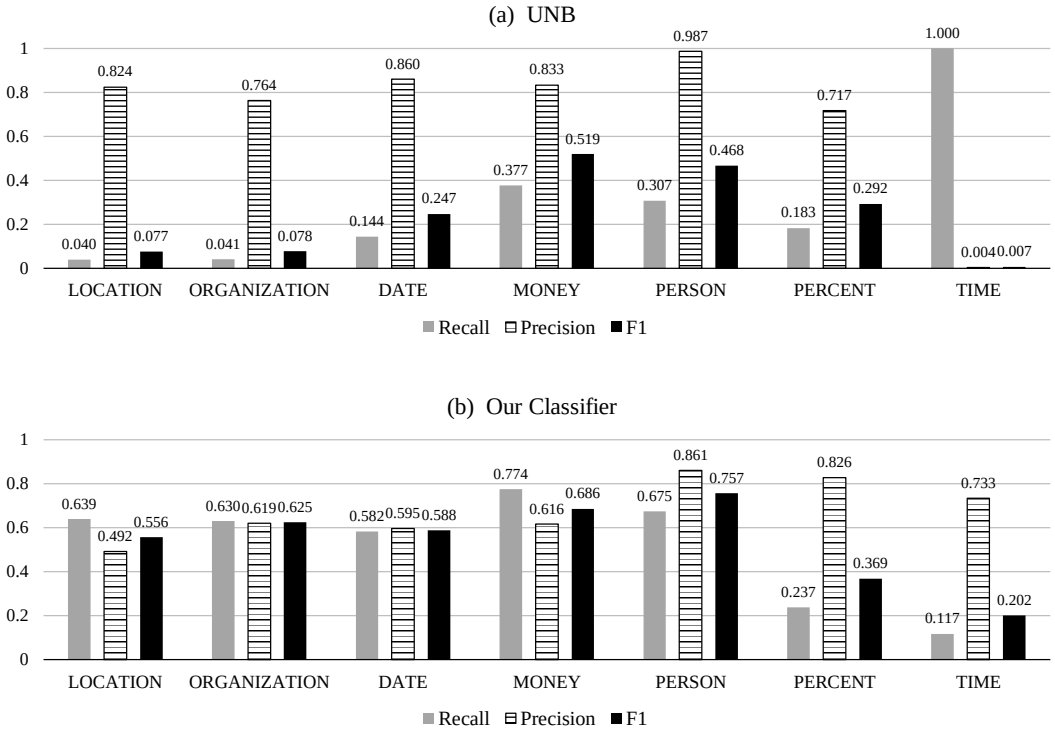

図 2　UNB と提案手法におけるクラスごとの再現率，適合率，F1 値（全クラス）.

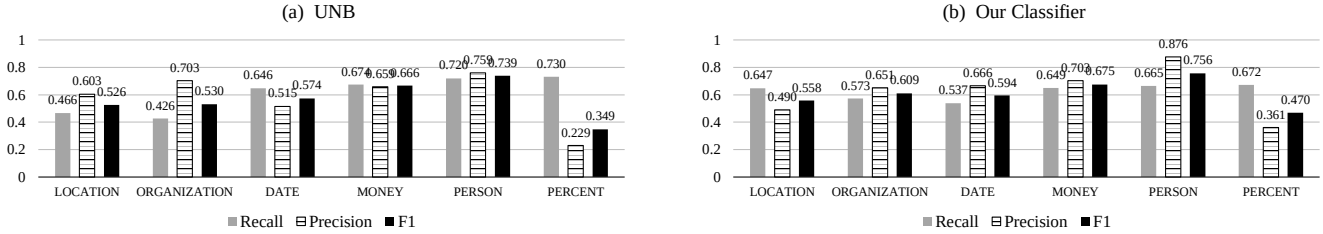

図 3　UNB と提案手法におけるクラスごとの再現率，適合率，F1 値（TIME 除外）.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.

[4] H. He, Y. Bai, E. A. Garcia, and S. Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proc. IJCNN'08*, pp. 1322–1328, 2008.

[5] I. Tomek. Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-6, No. 11, pp. 769–772, 1976.

[6] J. Zhang and I. Mani. kNN approach to unbalanced data distributions: A case study involving information extraction. In *Proc. the ICML'2003 Workshop on Learning from Imbalanced Datasets*, pp. 1–7, 2003.

[7] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proc. ICML'03*, pp. 616–623, 2003.

[8] C. Elkan. The foundations of cost-sensitive learning. In *Proc. IJCAI'01*, pp. 973–978, 2001.

[9] X. Chai, L. Deng, Q. Yang, and C. X. Ling. Test-cost sensitive naive bayes classification. In *Proc. ICDM'04*, pp. 51–58, 2004.

[10] X. Fang. Inference-based naive bayes: Turning naive bayes cost-sensitive. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 10, pp. 2302–2313, 2012.

[11] Y. Xiong, M. Ye, and C. Wu. Cancer classification with a cost-sensitive naive bayes stacking ensemble. *Computational and Mathematical Methods in Medicine*, pp. 1–12, 2021.

[12] W. Härdle, A. Werwatz, M. Müller, and S. Sperlich. *Nonparametric and Semiparametric Models*. Springer, 2004.

[13] 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.

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

[15] 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. 1433–1440, 2008.

[16] 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.

[17] M. Kikuchi, M. Yoshida, K. Umemura, and T. Ozono. Feature selective likelihood ratio estimator for low- and zero-frequency N-grams. In *Proc. ICAICTA 2021*, pp. 1–6, 2021.

[18] R. Storn and K. Price. Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, Vol. 11, No. 4, pp. 341–359, 1997.

[19] 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.

表 8　CNB と NNB の分類結果（全クラス）.

分類器	Macro 再現率	Macro 適合率	Macro F1 値	Micro 正解率
NB	♣ .546	.510	.483	.529
CNB	.477	.571	● .489	● .618
CNB（$p(c)=1$）	.462	.471	.431	.514
NNB	.467	.461	.441	.552
UNB	.299	♣ .713	.241	.160
提案手法	● .522	● .678	♣ .540	♣ .626

表 9　CNB と NNB の分類結果（TIME 除外）.

分類器	Macro 再現率	Macro 適合率	Macro F1 値	Micro 正解率
NB	.609	.574	.571	.582
CNB	.556	♣ .669	.571	♣ .620
CNB（$p(c)=1$）	.597	.567	.575	.597
NNB	.597	.588	● .588	.611
UNB	● .610	.578	.564	.568
提案手法	♣ .624	● .624	♣ .610	● .617

表 10　CNB と NNB におけるクラスごとの再現率，適合率，F1 値．各クラスに対する F1 値の最大値を記号 ♣，次いで大きい値を記号 ● で強調している．

分類器	LOCATION 再現率	LOCATION 適合率	LOCATION F1 値	ORGANIZATION 再現率	ORGANIZATION 適合率	ORGANIZATION F1 値	DATE 再現率	DATE 適合率	DATE F1 値	MONEY 再現率	MONEY 適合率	MONEY F1 値
NB	.416	.601	.491	.439	.662	.528	.579	.526	.551	.666	.642	.654
CNB	.519	.586	● .550	.718	.570	♣ .636	.421	.716	.530	.726	.604	● .659
CNB（$p(c)=1$）	.460	.634	.533	.451	.687	.545	.488	.634	.551	.703	.605	.650
NNB	.489	.616	.545	.540	.652	.591	.491	.643	● .557	.727	.601	.658
UNB	.040	.824	.077	.041	.764	.078	.144	.860	.247	.377	.833	.519
提案手法	.639	.492	♣ .556	.630	.619	● .625	.582	.595	♣ .588	.774	.616	♣ .686

分類器	PERSON 再現率	PERSON 適合率	PERSON F1 値	PERCENT 再現率	PERCENT 適合率	PERCENT F1 値	TIME 再現率	TIME 適合率	TIME F1 値
NB	.637	.798	.708	.598	.331	♣ .427	.489	.011	● .021
CNB	.749	.692	.720	.204	.831	.327	0	0	0
CNB（$p(c)=1$）	.727	.734	● .730	0	0	0	.404	.005	.010
NNB	.738	.711	.724	0	0	0	.287	.006	.011
UNB	.307	.987	.468	.183	.717	.292	1	.004	.007
提案手法	.675	.861	♣ .757	.237	.826	● .369	.117	.733	♣ .202

付　　録

補集合を用いた他の分類器である，Complement Naive Bayes 分類器（CNB），Negation Naive Bayes 分類器（NNB）と提案手法との比較結果を述べる．訓練データとして実データを用い，5.2 節に示す手順に従い実験を行った．

CNB　補集合を用いた有名な分類器である CNB [7] は

$$\widehat{c}(y) = \arg\max_{c \in C} p(c) \prod_{k=1}^{n} \frac{1}{p(w_k \mid \bar{c})}$$

と定義される．$p(w_k \mid \bar{c})$ の推定にはラプラススムージングを利用する．なお CNB の原著論文では，$p(c)$ が推定に与える影響は $p(w_k \mid \bar{c})$ よりも小さいと判断し，分類問題を解く際には $p(c)$ を無視して $p(w_k \mid \bar{c})$ のみを推定している．そこで $p(c)=1$ と固定した CNB も比較対象に加える．

NNB　CNB はヒューリスティックな解法であり，事後確率最大化の式から導出できない．そこで，事後確率最大化の式から導出でき，かつ補集合を用いた分類器である NNB [1] が提案された．NNB は

$$\widehat{c}(y) = \arg\max_{c \in C} \frac{1}{1-p(c)} \prod_{k=1}^{n} \frac{1}{p(w_k \mid \bar{c})}$$

と定義される．$p(w_k \mid \bar{c})$ の推定にはラプラススムージングを利用する．

CNB と NNB の分類結果を表 8 と表 9 に示す．比較のために NB，UNB，提案手法の分類結果も表中に再掲した．各指標の最大値を記号 ♣，次いで大きい値を記号 ● で強調している．これらの表から，CNB および NNB と比較しても提案手法の性能は優れており，正則化パラメータの最適化基準とした Macro-F1 は提案手法が最良なことが分かる．

また，各クラスに着目した分類結果を表 10 に示す．この表から分かるように，提案手法は 7 クラス中 5 クラスで最大の F1 値を達成し，他の 2 クラスでも 2 番目に大きい F1 値を持つ．その一方で CNB と NNB は，少数派クラスである PERCENT や TIME へインスタンスを一つも分類できないケースがあった．この場合，適合率と F1 値が計算不能となるため，CNB および NNB の該当する適合率と F1 値をゼロとした．提案手法は確度の高いインスタンスのみを TIME へ分類する傾向がある．対して CNB と提案手法を除く分類器では，TIME の適合率と F1 値が小さく，多数のインスタンスを少数派クラスの TIME へと誤分類する様子が示された．