DEIM 2020 F8-4(day2 p55)

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

1. 観測頻度に基づく尤度比の推定

■ 尤度比の定義と推定例

尤度比：確率分布の比で表される統計的尺度

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

確率密度 $p_*(x)$, $*\in \{\mathrm{de}, \mathrm{nu}\}$

N-gram $x = \langle a_1, a_2, \cdots, a_N \rangle$,　a_k は単語等の離散値

● 素朴な推定法：確率分布を相対頻度で求め、その比を取る

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

$f_*(x)$ は密度 $p_*(x)$ の確率分布から観測された x の頻度

Table1　尤度比の推定例

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

● 素朴な推定法の問題点

1）低頻度の問題　：$f_*(x)$ が低いとき、推定値が不当に大きくなることがある（x_{B} と x_{C}）

2）ゼロ頻度の問題：$f_*(x)$ が0のとき、有効な推定値を算出できない（w_{D}）

● 先行研究：正則化を導入した手法 [2]

uLSIF [1]：最小二乗アプローチによって、確率密度推定を介さずに尤度比を直接推定

uLSIFの枠組みに従い、低頻度の問題を緩和 [2]

⇒ 元の枠組みから基底関数を変更し、離散要素の扱いを可能とした

$$\hat{r}(x) = \frac{\hat{p}_{\mathrm{nu}}(x)}{\hat{p}_{\mathrm{de}}(x) + \lambda}$$

◎ 正則化パラメータ λ により、$f_{\mathrm{de}}(x)$ を考慮し $r(x)$ を低めに推定

× ゼロ頻度の問題に対処していない（w_{D}）

2. 提案する推定量

■ 低頻度に加えてゼロ頻度にも対処

● x を a_k 単位に分解し、a_k に対する尤度比の積を取る

$$r_{\mathrm{R}}(x) = \prod_{k=1}^{N} \hat{r}(a_k) = \hat{r}(a_1) \times \hat{r}(a_2) \cdots \times \hat{r}(a_N)$$

◎ $r(a_k)$ の推定に手法 [2] を用いると、低頻度の問題を緩和できる

◎ 要素 x がなくても、a_k が出現していれば計算できる

× （暗黙に）要素間の出現に独立性を仮定

⇒ **だが、この仮定は実際には成り立たないことが多い**

● 特徴重み付け法を応用（**提案手法**）

$$r_{\mathrm{ours}}(x) = \prod_{k=1}^{N} \hat{r}(a_k)^{W_k} = \hat{r}(a_1)^{W_1} \times \hat{r}(a_2)^{W_2} \cdots \times \hat{r}(a_N)^{W_N}$$

特徴変数を A_k とし、変数が取る値（単語など）を a_k とする

◎ **A_k に対する重み W_k が独立性の仮定を緩和**

● Correlation-based Feature Weighting (CFW) filter [3]

ナイーブベイズ分類器のために開発された、相関ベースの重み計算法

分類への影響小な A_k には0に近い重み、影響大な A_k には1に近い重み

[1] T. Kanamori *et al.*, A least-squares approach to direct importance estimation. JMLR, Vol. 10, pp. 1391-1445, July 2009.

[2] 菊地 *et al.*, 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.

3. 実験設定

■ 固有表現（地名・人名）の左N-gramを予測

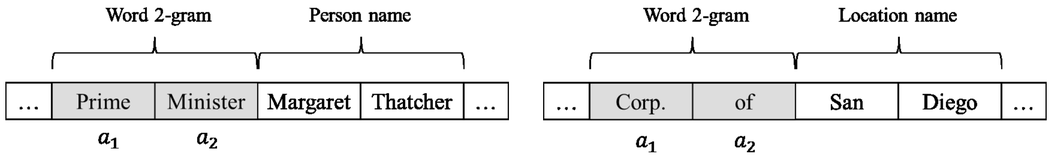

Fig. 1　固有表現の左文脈（N=2）

● データセット

1987 Wall Street Journal Corpus へ固有表現タグを付与

記事を訓練・開発・評価データへと分配（開発・評価データは1000記事に固定）

● 実験条件：3条件から一つ選び、その組み合わせで実験

条件① N-gramの次数N　　：2, 4, 6, 8, 10

条件② 訓練データのサイズ：2500, 5000, 7500, 10000記事

条件③ 固有表現：地名, 人名

● 実験手順

（1） 訓練データの N-gram に対して頻度をカウント

（訓練データを用いて重み計算、開発データを用いて正則化パラメータ決定）

（2） 評価データにある N-gram x に対して次の尤度比を推定

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

$p_{\mathrm{ne}}(x)$ ：x が固有表現（地名 or 人名）の左に出現する確率

$p_{\mathrm{tr}}(x)$ ：x が訓練データの任意位置に出現する確率

（3） x を推定値の降順に並べ、ランク—再現率曲線を描く

● 提案手法【正則化+重み付け】に対する比較手法

【ベースライン】確率分布を相対頻度で求め、その比を取る素朴な手法

【重み付け】　重み付けのみ使用、正則化パラメータ $\lambda = 0$

【正則化】　正則化のみ使用、重み $W_k = 1$

4. 実験結果（論文から一部抜粋）

■ ランク—再現率曲線

原点と曲線上の点を結ぶ直線の傾きがそのランクまでの適合率に比例

$$\text{再現率} = \frac{\text{そのランクまでの 正解数}}{\text{テストデータに含まれる正解数}} \qquad \text{適合率} = \frac{\text{そのランクまでの正解数}}{\text{正誤判定したランク}}$$

※ 正解数は、評価データにおいて固有表現の左に現れた x の種類数

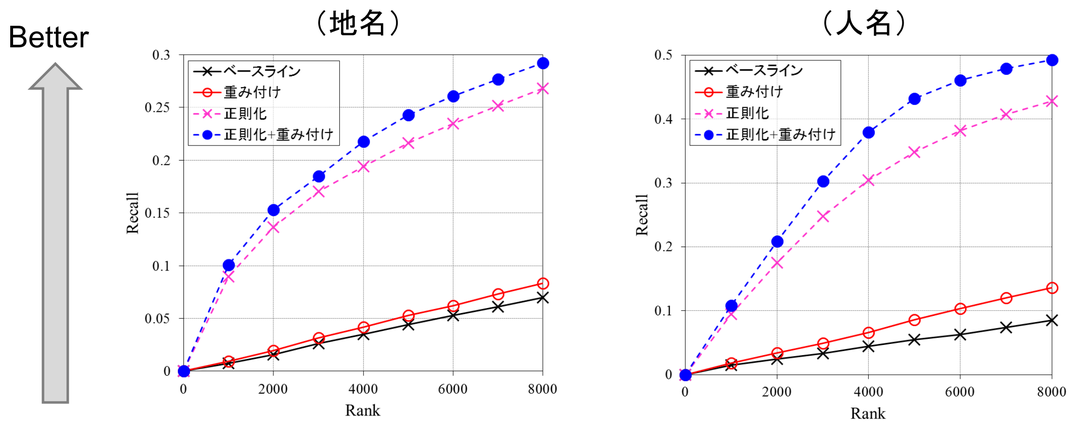

Fig. 2　N=10、訓練データ10000記事

● 結果：N が大きいときに、提案手法は極めて有効

N が大きいほど独立性仮定の影響が大きく、正則化のみの手法との差が明確

■ 今後の課題

課題1：A_k ではなく、a_k に対するより細かい重み付けで性能を更に向上させる

課題2：大きな N に対しても有効で安定した推定法を実現させる

課題3：提案手法が有効な実用タスクの探索

[3] L. Jiang *et al.*, A correlation-based feature weighting filter for naive Bayes. TKDE, Vol. 31, No. 2, pp. 201-213, 2018.