日本オペレーションズ・リサーチ学会
2021年 秋季研究発表会

確率的潜在意味解析における初期値決定方法の提案

05001400　中央大学大学院　* 寺澤 眞之介　TERASAWA Shinnosuke
05000907　東海大学　大竹 恒平　OTAKE Kohei
01405390　中央大学　生田目 崇　NAMATAME Takashi

1. 研究目的

確率的潜在意味解析 (pLSA) は次元圧縮手法の一つであり，行列に対して，行と列に共通な潜在クラスにより表現され，ソフトクラスタリングとしても利用される．pLSA は，高次元データに対応できるだけでなく，行と列の要素を同時にクラスタリングできるという特徴を持つため，文書データや ID-POS データなどではトピックモデルとして利用される [1]．しかし，pLSA の解は，求解手法の EM アルゴリズムの初期値に依存するため，解が安定しないことも知られている [2]．そこで本研究では，より効率的に最適な解を求めるための初期値設定方法を提案する．

2. 提案手法

2.1. pLSA の概要

pLSA は，行の要素 x と列の要素 y の背後には，共通する特徴となる潜在的な意味クラス z があると想定し，この潜在クラス z を確率的に計算する．x と y の共起確率 $P(x,y)$ を潜在クラス z を使って表現し，確率変数 $P(x \mid z)$，$P(y \mid z)$，$P(z)$ を最終的な結果として出力する．ここで共起確率 $P(x,y)$ は (1) 式のようにモデル化することができ，同時出現頻度を $N(x,y)$ とすると，対数尤度関数 L は (2) 式のようになる．

$$P(x,y) = \sum_z P(x \mid z) P(y \mid z) P(z) \tag{1}$$

$$L = \sum_x \sum_y N(x,y) \log P(x,y) \tag{2}$$

この対数尤度関数 L を最大にするような $P(x \mid z)$，$P(y \mid z)$，$P(z)$ を最尤推定する．最尤推定には EM アルゴリズムが用いられ，(3) 式を用いて z 以外の変数を固定して $P(z \mid x,y)$ を計算する E ステップと，(4)〜(6) 式を用いて $P(x \mid z)$，$P(y \mid z)$，$P(z)$ を算出する M ステップを，解が収束するまで繰り返す．

$$\begin{aligned} P(z \mid x,y) &= \frac{P(x,y,z)}{P(x,y)} \\ &= \frac{P(x \mid z) P(y \mid z) P(z)}{\sum_z P(x \mid z) P(y \mid z) P(z)} \end{aligned} \tag{3}$$

$$P(x \mid z) = \frac{\sum_y N(x,y) P(z \mid x,y)}{\sum_x \sum_y N(x,y) P(z \mid x,y)} \tag{4}$$

$$P(y \mid z) = \frac{\sum_x N(x,y) P(z \mid x,y)}{\sum_x \sum_y N(x,y) P(z \mid x,y)} \tag{5}$$

$$P(z) = \frac{\sum_x \sum_y N(x,y) P(z \mid x,y)}{\sum_x \sum_y \sum_z N(x,y) P(z \mid x,y)} \tag{6}$$

2.2. 提案する初期値設定方法

本研究では，pLSA でより効率的に安定した解を得るための初期値設定方法を提案する．pLSA は行と列の観点から同時に，近い要素同士を同じ潜在クラスとしてソフトクラスタリングする．そこで，要素間の距離を確率として，$P(x \mid z)$，$P(y \mid z)$ の初期値を設定する．本研究では，各要素の座標を算出する手法としてコレスポンデンス分析，近い要素をまとめる手法として k-means 法，各クラスタの重心座標と各要素の座標の類似度としてベクトルの内積，初期値として確率にする手法として softmax 関数を用いる．初期値作成の流れを以下に示す．

1. 共起行列に対してコレスポンデンス分析を実施し，正準相関係数，行の得点，列の得点を算出する．この際の次元数を，pLSA の潜在クラス数と揃える．
2. 各要素の行の得点，列の得点に対して，正準相関係数を掛け合わせ，各要素の座標を算出する．
3. 行の要素，列の要素ごとに座標に対して k-means 法でクラスタリングをする．この時のクラスタ数も pLSA の潜在クラス数と揃える．
4. 各クラスタの重心座標を算出し，各要素と各クラスタの重心との内積を算出する．
5. 算出した内積に対して，softmax 関数を用いて各クラスタにおける各要素の合計が 1 になるようにする．
6. 算出した値の行方向を $P(x \mid z)$ の初期値，列方向を $P(y \mid z)$ の初期値とし，EM アルゴリズムを実行する．

2.3. クラスタ番号の統一方法

初期値の違いによる pLSA の安定性を評価するためには，各結果のクラスタ番号を一致させる必要がある．そこで，1 回目の結果のクラスタ番号を固定し，2 回目以降のクラスタ番号の割当てを最適化問題を利用して行う．このとき，n 回目の結果において $P(x,y \mid z) = P(x \mid z) P(y \mid z)$ が成り立つ．1 回目の結果のクラスタ番号を固定しているため，1 回目の結果の潜在クラス i と，$n = l$ 回目の結果の潜在クラス j の近接度 sim_l を (7) 式のように定義する．

$$\begin{aligned} sim_l(i,j) = \sum_x \sum_y \{ & P(x,y \mid z = i, n = 1) \\ & \circ P(x,y \mid z = j, n = l) \} \end{aligned} \tag{7}$$

l 回目ですべての i,j に対して，$m \times m$ の行列に並べたものを 1 回目の結果と l 回目の結果から求められる近接度 SIM_l とする．さらに，割当問題の変数として，バイナリ変数 $t_{(i,j)}$ を用いて (8) 式のような V を作成する．ここで，$t_{(i,j)} = 1$ となるとき，1 回目の結果の潜在クラス i と l 回目の結果の潜在クラス j が同じクラスである．

$$V = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & t_{(1,1)} & \cdots & t_{(1,j)} & \cdots & t_{(1,m)} \\ \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 & t_{(i,1)} & \cdots & t_{(i,j)} & \cdots & t_{(i,m)} \\ \vdots & & \vdots & \ddots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 & t_{(m,1)} & \cdots & t_{(m,j)} & \cdots & t_{(m,m)} \end{pmatrix} \tag{8}$$

V の k 行目を V_k とすると割当問題を以下に示す．

$$\max \quad VV = \sum_{k=1}^{m} \left(V_k^T \times V_k\right) \circ SIM_l \tag{9}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} t_{(i,j)} + 1 = 2, \quad j = 1, \ldots, m \tag{10}$$

$$\sum_{j=1}^{m} t_{(i,j)} = 1 \quad i = 1, \ldots, m \tag{11}$$

$$t_{i,j} \in \{0, 1\} \quad i, j = 1, \ldots, m \tag{12}$$

これを各結果について行う．

2.4. pLSA の性能評価・比較

pLSA の性能を以下の 2 つの観点から評価する．

(i) EM アルゴリズムが収束するまでの反復回数

(ii) $P(x \mid z), P(y \mid z)$ の確率が上位の要素の共通度

(i) の観点では，pLSA は EM アルゴリズムを利用して最尤推定を行っているため，解が収束するまで反復を繰り返す．そこで反復回数を測定して，これを収束速度とみなす．

(ii) の共通度として Jaccard 係数すなわち，2 つの集合に含まれる要素の和集合に対する共通の要素の割合を用いる．

3. 評価実験

3.1. データ概要

評価実験として，あるスーパーマーケットの ID-POS データを利用する．対象データの概要を以下に示す．

- **分析対象期間** 2015 年 1 月 1 日〜2015 年 12 月 31 日
- **会員数** 134,058
- **商品カテゴリ数** 26　(ex. “野菜”，“果物”，...)
- **年間売上件数** 62,975,264 件

3.2. 実験の流れ

本研究では，行を「会員番号」，列を「商品カテゴリ」，要素を「購入数」とした共起行列を作成し，この共起行列を基に，提案手法の初期値を作成する．また，以下の 2 通りで設定した初期値を用い，pLSA を 10 回実行する．なお，潜在クラス数は 4 とした．

1. 実行する毎に初期値を全て乱数で発生させる．
2. 提案手法で作成した初期値を利用して，初期値を一部固定し，他の初期値を乱数で発生させる．

これらの結果を基に提案手法の性能の評価・比較を行う．

3.3. 実験結果

2 通りの初期値における各回の EM アルゴリズムの反復回数を表 1 に示す．

表 1: 2 通りの初期値における反復回数の比較 [回]

試行回	乱数	提案手法	試行回	乱数	提案手法
1	414	163	6	424	312
2	974	687	7	220	273
3	327	222	8	295	191
4	254	216	9	347	160
5	410	167	10	479	164
			平均	414.4	255.5

表 1 から，乱数で初期値を発生させた際の反復回数と，提案手法の反復回数を比較すると，提案手法の方が乱数での初期値設定時より，4 割程度削減できていることが分かる．

次に，2 通りの初期値における $P(x \mid z), P(y \mid z)$ の各クラスタの Jaccard 係数と，初期値を乱数で発生させた際の Jaccard 係数から作成した初期値を利用した際の Jaccard 係数を引いた値を表 2 に示す．本研究では 10 回の結果に対して Jaccard 係数を用いるため，積集合を 1〜10 回目のすべての組合せ 45 通りの共通要素数の合計，和集合を 1〜10 回目のすべての組合せ 45 通りで出現する要素とする．これを各クラスタについて行い，それらの平均を Jaccard 係数とする．2 通りの初期値で実行した pLSA の結果に対して Jaccard 係数を算出して比較する．また，上位の要素数を $P(x \mid z), P(y \mid z)$ のそれぞれについて，100 個と 5 個とする．

表 2: 2 通りの初期値における Jaccard 係数の比較

	乱数 (A)	本研究の方法 (B)	差分 $(B-A)$
$P(x\|z)_1$	0.2155	0.7814	0.5660
$P(x\|z)_2$	0.2111	0.3548	0.1437
$P(x\|z)_3$	0.1302	0.5369	0.4067
$P(x\|z)_4$	0.1662	0.5935	0.4273
$P(x\|z)$ 平均	0.1807	0.5666	0.3859
$P(y\|z)_1$	0.5095	0.6881	0.1786
$P(y\|z)_2$	0.3198	0.5149	0.1951
$P(y\|z)_3$	0.3107	0.5852	0.2745
$P(y\|z)_4$	0.2944	0.7206	0.4263
$P(y\|z)$ 平均	0.3586	0.6272	0.2686

表 2 より，2 通りの初期値における Jaccard 係数の差がすべて正であり，提案した初期値で実行した pLSA の解のほうが高いことが分かる．これらの結果から，本研究で作成した初期値で pLSA を実行すると，乱数で発生させた初期値で pLSA を実行した場合と比べて反復回数が削減でき，最終的な解も安定すると言える．

4. まとめと今後の課題

本研究では，pLSA に内在する初期値依存性の問題に対して，既存手法よりも効率的に初期値を設定する手法を提案した．さらに，提案した方法で作成した初期値と，従来の乱数で発生させた初期値で pLSA の性能を評価・比較を行った．比較の結果，提案手法で作成した初期値による pLSA の解の方が性能が優れていることが示された．しかし，本研究では初期値作成の段階で k-means 法を利用していることから，pLSA の解が k-means 法で利用する初期値に依存する可能性があるため，初期値作成の際に距離が近い要素をまとめる手法を検討する必要がある．

参考文献

[1] 内山俊郎：“情報理論的クラスタリングを用いた確率的潜在意味解析の性能向上”，電子情報通信学会論文誌 D, Vol. J100-D, No. 3, pp.419–426 (2017)

[2] 野守耕爾，神津友武：“三位一体アプローチによるテキストデータモデリング法の開発-宿泊施設の口コミデータを用いた評価推論モデルの構築-”，人工知能学会全国大会論文集 (第 28 回)，JSAI2014, 1L2OS17a1 (2014)