LDA と pLSA の推定結果の比較とその考察

株式会社豊田自動織機/産業技術総合研究所　原田奈弥*
東北大学　石垣司
産業技術総合研究所　本村陽一

１．はじめに

トピック分析に用いられる手法として、pLSA や LDA が多く用いられる。どちらの手法も、観測された word に対して潜在効果を仮定し、同じ潜在効果を持つ word をクラスタリングする手法だが、データに仮定するモデルが異なる。pLSA は、word 同士の共起確率とベイズの公式から潜在クラスを推定する。一方 LDA は、潜在効果に仮定したディリクレ分布を推定している。

２．問題

本研究は、pLSA と LDA の、実用上での比較を行うことを目的としている。

縦断データでトピック分析を行う場合、2 次元の観測データの一方の次元に対して、もう一方の次元の観測値が非常に大きいことが考えらえる。例えば、購買日毎に購買される商品を示す PoS データがあるとき、店舗などで販売する商品は数が決まっている一方、購買日は販売を行う限り増え続ける。x_i と y_j を観測したときに、pLSA の潜在クラス u_l の尤度関数は、

$$L = \sum_i \sum_j \left[N_{ij} \log\{P(x_i) \sum_l P(y_j \mid u_l) P(u_l \mid x_i)\} \right] \quad (1)$$

N_{ij} は、観測値 x_i と y_j の共起行列について、x_i と y_j が同時に観測された回数を示す。ある潜在クラス u_l に属する x_i と y_j の数が大きく異なることが、縦断データでは考えられる。このような構造のデータでトピック分析を行った場合、pLSA と LDA の結果がどのように異なるかを検証することは、これらの手法の実用上の価値がある。

３．検証

上記のような潜在効果の構造を持った仮想データを作り、pLSA と LDA の推定結果について、比較を行う。これらの２つの手法の結果の違いを生じさせる要因について、モデルの違いから考察を行う。

---

*135-0064 東京都江東区青梅 2-4-7 産業技術総合研究所 臨海副都心センター別館 9 階
name.harada@aist.go.jp