疑似ラベルを用いた潜在的ディリクレ配分法への取り組み

鈴木聡子　(指導教員：小林一郎)

1　はじめに

近年，文書の潜在情報であるトピックを考慮したトピックモデルが文書要約や文書分類に利用されている．潜在ディリクレ配分法 (LDA)[1] に基づいて提案された Labeled LDA(L-LDA)[2] は，人によって文書に付けられたタグを，その文書の意味内容を表すものと捉え，潜在トピック抽出における教師信号として利用することを考えたモデルであり，複数のタグ付き文書に対しての LDA を上回る性能を示すと知られている．しかし実際は，世の中のほとんどの文書にはタグが付与されておらず，L-LDA の使用される範囲は限られている．そこで本研究では，文書集合からタグの代わりとなる疑似ラベルを作成し，全ての文書に対して L-LDA が有用になることを目的とする．

2　Labeled LDA

L-LDA は，LDA におけるトピック分布を推定する過程で，文書に付与されたタグの情報を考慮したモデルとなっている．図 1 に L-LDA のグラフィカルモデルを示す．

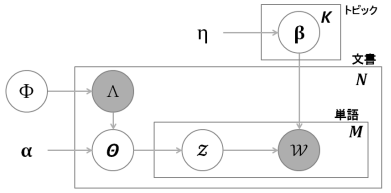

図 1: L-LDA のグラフィカルモデル

L-LDA と LDA との違いは，ラベル（文書に与えられているタグ）の情報が，θ を推定する際に影響を与えているという点である．

まず，文書ごとに付与されているタグの情報から，文書ラベル $\Lambda^{(d)}$ を生成する．

$$\Lambda^{(d)} = (l_1, \cdots, l_K) \quad l_k \in \{0, 1\} \tag{1}$$

K は文書群に含まれる重複の無いラベルの個数であり，文書ごとにラベルの有無の情報を 1 または 0 の 2 値で与える．次に文書におけるラベルのベクトルを定義する．

$$\lambda^{(d)} = \{k | \Lambda_k^{(d)} = 1\} \tag{2}$$

$\lambda^{(d)}$ は，文書 d に付与されているラベル番号である．そして，文書ごとに射影行列を生成する．

$$L_{ij}^{(d)} = \begin{cases} 1 & if\ \lambda_i^{(d)} = j \\ 0 & otherwise. \end{cases} \tag{3}$$

生成した射影行列と設定したハイパーパラメータ α から，文書ごとに新しいパラメータ $\alpha^{(d)}$ を生成する．ラベルの情報により制限された $\alpha^{(d)}$ から，トピック分布 θ を求める．他の過程は，LDA と同様である．

3　疑似ラベル生成

本研究では，以下の文書の表層的な 2 つの情報からラベルの代わりとなる疑似ラベルを生成する．

3.1　単語の共起情報に基づく疑似ラベル生成

まず，疑似ラベルを生成するために候補となる単語として，文書ごとに TF-IDF の高いものを抽出する．そして，抽出した単語の出現頻度を全ての文書において求める．ここで文書頻度が 1 である単語は，特定の文書においてのみ現れている単語であるため，疑似ラベル候補の中から消去する．文書の潜在的意味の一貫性は単語の共起関係に関連があるという Newman らの研究 [3] を参考に，生成するラベルにそのような意味情報が含まれるように，抽出した単語の自己相互情報量（PMI）を求め，共起関係の強い単語群を 1 つのグループとする．そのグループで 1 つの疑似ラベルを表すとする．

また，共起情報によるクラスタリングで作られたラベルの他に，PMI の値は低いが出現頻度は高い単語も，ラベルとして採用する．

3.2　文書の類似度に基づく疑似ラベル生成

文書分類を行い，類似する文書に同じラベルを付与する．ここで，単純な Leader-Follower 法と Crouch 法の 2 つの方法 [5] において疑似ラベルを生成する．この 2 つの方法は，分類の重複を許すアルゴリズムとなっており，1 文書に対し複数のラベルを生成することが可能である．

Leader-Follower 法

文書ベクトルをもとに類似度を計算し，クラスタリングを行う．ここで用いるのは，単純な Leader-Follower 法である．本研究で用いたアルゴリズムの概要を以下で説明する．

1. 文書をクラスタに併合するための閾値を設定する．
2. 1 文書ずつ読む．全ての文書が読み終わったら処理を終了する．
3. 読み込んだ 1 文書と，その時点で存在する全てのクラスタとの類似度を計算する．
4. 閾値以上の類似度をもつクラスタにその文書を併合し，クラスタの語の重みの値を更新する．その文書との類似度が，どのクラスタにおいても閾値を超えない場合は，新しいクラスタとして生成する．
5. 手順 2 に戻る．

これによって生成されたクラスタについて，同じクラスタに含まれている文書に同じ疑似ラベルを振る．ここで，クラスタを構成する文書数が 1 の場合，疑似ラベルを振らないとする．

Crouch 法

この手法は Leader-Follower 法を拡張したものであり，クラスタの設定とクラスタへの文書の割り当てを 2 段階の処理によって行うことが特徴である．

4　実験

タグ付けされていない文書集合に疑似ラベルを付与し，文書分類の課題を通じて LDA との比較を行う．

4.1 実験仕様

使用するデータは，20 Newsgroups[1]の内 10 カテゴリの中からそれぞれ 100 文書をランダムに選んだ，合計 1000 文書を用いる．それらは，ストップワードを除いた後，ステミング処理を施す．提案手法の実験は，2 つのパターン（1 : 単語の共起情報により疑似ラベルを生成した場合，2 : 文書の類似度から疑似ラベルを生成した場合）において行う．

パターン 1 では，抽出する単語数は各文書ごとに TF-IDF の値が上位 30 単語とした．ここで，予備実験より PMI の閾値はラベルが複数できる範囲 [4.5,6.2] において実験を行うものとした．パターン 2 では，Leader-Follower 法と Crouch 法の 2 つの方法で疑似ラベルを生成した．両方とも，閾値を [0.1,0.9] と [0.03,0.1] において実験を行った．全ての実験で，L-LDA に与えるハイパーパラメータの値は，α=0.1，η=0.1 とした．

比較対象である LDA では，まずトピックを設定するために予備実験を行った．トピック数ごとに，イテレーションの中での最も低いパープレキシティの平均を 10 回の試行から求めた．その結果，トピック数が 16 で，平均が最小となったため，実験ではトピック数を 16 と設定した．LDA において与えるパラメータの値は，提案手法と同じく α=0.1，η=0.1 とした．

4.2 評価手法

文書のトピック分布 θ から，各文書のトピックで構成されるベクトルを作り，k-means 法により，20 Newsgroups の対象とした 10 カテゴリのグループに文書を分類した際の精度を見ることで提案手法の評価を行う．評価手法には，文献 [4] で用いられている評価手法を採用し，式（4）に示される相互情報量を利用した．

$$MI(L,A) = \sum_{l_i \in L, \alpha_j \in A} P(l_i,\alpha_j) \cdot \log_2 \frac{P(l_i,\alpha_j)}{P(l_i)P(\alpha_j)} \quad (4)$$

$L = \{l_1, l_2, \cdots, l_k\}$ は，k-means 法により分類された文書ラベルの集合であり，$A = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ は分類された文書の正解ラベルである．また，$P(l_i)$ は分類により l_j にラベル付けされる確率，$P(\alpha_j)$ は正解データにおいて α_j である確率，$P(l_i,\alpha_j)$ はこれら 2 つが同時に起こる確率である．

ここで，相互情報量を [0,1] の値で得るために式（5）により正規化を行う．

$$\widehat{MI} = \frac{MI(L,A)}{MI(A,A)} \quad (5)$$

4.3 実験結果

k-means 法を用いた分類をそれぞれの手法につき 10 回行い，$\widehat{MI}$ の平均を求めた．実験結果を図 2〜4 に示す． 全てのグラフに関して，横軸は閾値，縦軸は評価値を示す．なお，比較のため LDA の評価結果もグラフに示す．図 2 から分かるように，パターン 1 ではどの閾値においても LDA よりも低い評価値となった．パターン 2 においては，[0.1,0.9] では 2 つの手法は類似した形のグラフとなっているが，Leader-Follower 法を用いた場合の方が，値の差が大きい．また，[0.03,0.1] では Leader-Follower 法を用いた場合，閾値が小さいときは良い結果を示しているが，閾値が大きくなるにつれて $\widehat{MI}$ は減少していく．一方，Crouch 法を用いた場合，安定して LDA に近い値となっており，一部では上回っている．

[1]http://qwone.com/ jason/20Newsgroups/

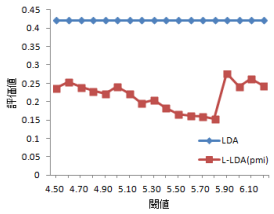

図 2: $\widehat{MI}$(パターン 1)

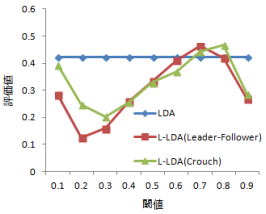

図 3: $\widehat{MI}$(パターン 2a)

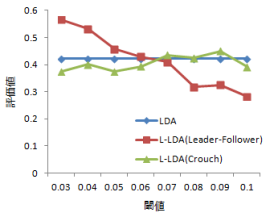

図 4: $\widehat{MI}$(パターン 2b)

5 考察

実験結果より，単語の共起情報から疑似ラベルを生成した場合では，どの閾値においても LDA と比べ精度が上がらなかった．これは，単語の共起情報のみでは文書集合の全てのトピックについて反映しきれていない，また，閾値を正しく設定することが困難なためと考えられる．文書の類似度から疑似ラベルを生成した場合，Leader-Follower 法の方が評価結果の値が閾値によって大きく変化した．その原因としては，Leader-Follower 法ではクラスタを生成する際に文書を順に読み込むことから，文書の処理の順番にクラスタの生成が依存するためだと考えられる．

6 おわりに

本研究では，単語の共起情報と文書の類似度の 2 つの表層的な情報から疑似ラベルを生成した．生成した疑似ラベルを用いて実験を行い，LDA との精度の比較，評価を行った．結果，閾値を変えることによって，一部 LDA よりも良い精度を得ることができた．

今後は，両手法の精度の向上を検討し，両手法により生成された疑似ラベルを併せて利用した場合について実験を行う．

参考文献

[1] D. M. Blei, A. Y. Ng, M. I. Jordan : Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
[2] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning : Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. EMNLP2009, pp. 248-256, 2009.
[3] Newman, David and Lau, Jey Han and Grieser, Karl and Baldwin, Timothy : Human Language Technologies, NAACL2010,pp. 100–108,Los Angeles, California, 2010.
[4] Gunes Erkan : Language Model-Based Document Clustering Using Random Walks, Association for Computational Linguistics,pp.479-456,2006.
[5] 岸田和明:文書クラスタリングの技法, Library and Information Science,No.49, 2003.