潜在情報を加味した教師データによるグラフを用いた文書分類

江里口 瑛子　(指導教員：小林 一郎)

1　序論

機械学習手法には複数あり，その中でグラフ構造に基づく半教師あり学習 (Graph-based Semi-Supervised Learning: GBSSL) 法は，SVM などの学習法と比べてより有効な手法である [1]．GBSSL 法の精度は，グラフ構成の仕方や，教師データ (ラベルありデータ) の選出の仕方によって左右される．

本研究は，グラフ構成に関連して，全カテゴリで一律に最適パラメータを決定するよりも各カテゴリ毎に決定する方が優れていることを示し，教師データの選出に関連しては新しい手法を提示し，機械学習の精度向上を目指す．まず，文書間の類似度に関して，これまで用いられてきた表層情報に基づく類似度に加えて，新たに，文書間の潜在情報に基づく類似度を加味したものを採用する．次いで，この両情報を総合した類似度に基づくグラフを作成し，TopicRank[2] 法を用いて，質の高い教師データを選出する．その際，グラフのノードとしては，単文ではなく，文書 (文集合) を採用する．この新しい手法を，マルチラベルを有するテキストのカテゴリ分類に適用して実験を行い，その有効性を検討し評価する．

2　グラフに基づく文書分類手法

2.1　グラフ構成

グラフは，ノード間の類似度を重みとする重み付き無向グラフ $G=(V,E)$ を用いる．ここで V と E は，それぞれグラフのノード集合と辺集合を表す．グラフ G は隣接行列 $\boldsymbol{W}$ の形で表現することができ，$w_{ij}\in\boldsymbol{W}$ はノード i，ノード j 間の類似度を表すとする．特に，GBSSL 法で用いる類似度は，ノード i の k-近傍点集合 $K(i)$ からなるものとし，$w_{ij}=sim(\boldsymbol{x}_i,\boldsymbol{x}_j)\delta(j\in K(i))$ とする．$\delta(z)$ は z が真ならば 1，偽ならば 0 とする．

2.2　グラフにおける類似度

一般にテキストデータを対象にしたグラフ構成において，ノード間の類似度としては単語の出現頻度に着目した *tfidf* のコサイン類似度が用いられる．これは文書の表層情報に基づく類似度を表している．本研究では，この従来の類似度に，新たに，文書の持つ潜在情報に基づいた類似度を $\alpha(0\leq\alpha\leq 1)$ の割合で付加し，これらを合算してノード間の類似度とする (式 (1))．式 (1) における P と Q は，それぞれ文書 S と文書 T のトピック分布を表す．トピック分布の推定法には，Latent Dirichlet Allocation(LDA)[3] を用い，トピック分布の類似度指標には，式 (2) によって Jensen-Shannon ダイバージェンス (D_{JS}) を類似度に変換したものを用いる．

$$sim(S,T)\equiv\alpha * sim_{JS}(P,Q) + (1-\alpha)*sim_{\cos}(tfidf(S),tfidf(T)) \quad (1)$$

$$sim_{JS}(P,Q)\equiv 1-D_{JS}(P,Q) \quad (2)$$

2.3　教師データ (ラベルありデータ) の選出

教師データの選出は，北島ら [2] の TopicRank 法を採用して行う．TopicRank 法とは，ノードを文とし，文間の潜在情報に基づく類似度で構成されたグラフに対して，式 (3) を用いて文の重要度を算出し，順位付けを行う手法である．ここで，d は制動係数 (damping factor) である．

本研究では，この方法を，グラフのノードを文から文書 (文の集合) に差し替えて用いる．データの選出にあたっては，文書のトピック分布を考慮した，教師データのみをノードにもつグラフをカテゴリ毎に作成し，TopicRank 法を用いて，スコアが高いデータから順に，GBSSL 法で用いる教師データとしていく．なお，式 (3) における N は対象文書群の総文書数，$adj[u]$ は文書 u の隣接ノード集合を表す．

$$r(u)=d\sum_{v\in adj[u]}\frac{sim(u,v)}{\sum_{z\in adj[v]}sim(z,v)}p(u)+\frac{1-d}{N} \quad (3)$$

2.4　ラベル伝搬法

ここでは学習法として，ラベル伝搬法 [4] を採用する．これは，カテゴリラベル未知のノードについて予測を行う方法であり，「グラフ上において，辺で繋がるノード同士は同じカテゴリに属す」という仮定に基づいている．この仮定に基づき類似度行列を $\boldsymbol{W}$，ノード数を n 個 (このうち，教師データ数は l 個) とする．n 個のノードに対する予測値 $\boldsymbol{f}$ は，以下の最適化問題の目的関数 (式 (4)) の解 (式 (5)) として求まる．$\boldsymbol{L}(\equiv\boldsymbol{D}-\boldsymbol{W})$ はラプラシアン行列と呼ばれ，対角行列 $\boldsymbol{D}$ は $\boldsymbol{W}$ の各行 (又は列) の和を対角成分に持つ行列である．

$$\begin{aligned} J(\boldsymbol{f}) &= \sum_{i=1}^{l}(y^{(i)}-f^{(i)})^2+\lambda\sum_{i<j}w^{(i,j)}(f^{(i)}-f^{(j)})^2 \\ &= ||\boldsymbol{y}-\boldsymbol{f}||_2^2+\lambda\boldsymbol{f}^T\boldsymbol{L}\boldsymbol{f} \quad (4) \\ \boldsymbol{f} &= (\boldsymbol{I}+\lambda\boldsymbol{L})^{-1}\boldsymbol{y} \quad (5) \end{aligned}$$

3　実験

3.1　実験仕様

実験の対象データとしては，Reuters-21578[1] (Reuters) を用いる．このデータセットから，“ModApte”分割に従って，本文とタイトルのみからなる記事データを抽出し，この全データに対してストップワードの除去とステミング処理を行う．その後，GBSSL 法を用いて文書分類を行っている Subramanya ら [1] の実験仕様に合わせ，10 種のカテゴリ **earn**，**acq**，**money-fx**，**grain**，**crude**，**trade**，**interest**，**ship**，**wheat**，**corn** に対する分類精度を求める．その際，マルチラベルを有する Reuters の記事データに対し，各カテゴリ毎に one-versus-rest 法を適用し，ある一定の閾値以上のカテゴリラベルを文書に付与するラベルとして採用する．この付与の成功度を表す指標が，指標 PRBEP であり，これを分類精度として採用する．これは

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/　Reuters は 135 のトピックカテゴリからなる Reuters newswire の英文記事を集めたデータセットである．

Precision(適合率) と *Recall*(再現率) が一致するときの値である．

データセットは，テストデータ (ラベルなしデータ)$u = 3299$ 個を共通とし，これに教師データ (ラベルありデータ)$l = 20$ 個を加えたものを 10 セット用意する．ちなみに，各データセットに含まれるデータ総数は $n = 3319$ 個である．教師データのカテゴリは，先述の 10 種のカテゴリに，新たなカテゴリ **others**(先述の 10 種のカテゴリに属さない文書を全て含む) を加えた全 11 種とする．また，データセットに加える教師データ l 個は上記 11 種のカテゴリからランダムに選択する．その際，各 11 種のカテゴリの教師データが少なくとも 1 個は含まれるようにする．

LDA 法による潜在トピック分布の推定には，ギブスサンプリングを用い，その反復回数は 200 回とする．トピック数はパープレキシティの値を算出し，その 10 回平均の値で決定する．他方，TopicRank 法で用いるグラフは，ノード数 $|V|$ =(カテゴリ毎の教師データの総数)，辺数 $E = |V \times V|$ の完全グラフとする．式 (1) におけるパラメータ α は，0.0 から 1.0 まで 0.1 刻み毎の値を与え，式 (3) における制動係数 d は 0.85 とする．

以上の設定下で，カテゴリ毎に各文書の TopicRank スコアを算出する．次いで，テストデータに加える教師データのカテゴリ数にしたがって，スコアの高い教師データから順にデータセットに加えていく．$\alpha = 0$ のときは文書の表層情報のみを扱い，類似度が一意的に決まるので推定を行う必要がなく，スコアは 1 回のみ算出する．他方，$\alpha \neq 0$ のときは文書の潜在トピック分布の推定を行わなければならず，類似度が一意的に決まらない．このため，スコアの 5 回平均の値をその文書のスコアとする．

ラベル伝搬法で用いた類似度グラフのノード数は $|V| = n(= 3319)$ である．ノード間の類似度は，コサイン類似度 (式 (1) で $\alpha = 0$ の時に相当) とし，表層情報のみからなるものとする．k-近傍グラフの大きさのパラメータ k は $\{2, 10, 50, 100, 250, 500, 1000, 2000, n\}$，ラベル伝搬法のパラメータ λ は $\{1, 0.1, 0.01, 1e{-}4, 1e{-}8\}$ の範囲を動かす．この設定に基づいて，以下の実験を行い，各 α 毎に各カテゴリ毎の PRBEP を求める．

実験 1 全カテゴリで共通の (一律の) 最適パラメータ (k, λ) の組を用意し，データセット 10 セットに対して文書分類を行う．最適パラメータは江里口ら [5] を採用する．

実験 2 各カテゴリに対する最適パラメータ (k, λ) の組をそれぞれ用意し，同様に 10 セットに対して文書分類を行う．最適パラメータは江里口ら [6] を採用する．

3.2 実験結果

実験 1, 2 の結果を図 1, 2 に示す．各図は 10 回の試行の全カテゴリ PRBEP のマクロ平均値とその標準偏差を示している．$\alpha = 0$ のときは表層情報のみ，$\alpha = 1.0$ のときは潜在情報のみを用いた場合であり，それ以外 $(0 < \alpha < 1)$ のときは，潜在情報と表層情報を一定の割合 $(\alpha : (1-\alpha))$ で混合した場合の結果を示している．

図 1，2 の点の変移を概観すると，共通する特徴が見られる．$0 < \alpha \leq 1$ の点は全て $\alpha = 0$ のときの点の上方にあり，$\alpha = 0.2, 0.6, 0.9$ で極大 ($\alpha = 0.6$ で最大) となっている．さらに $\alpha = 0$〜0.2 では，PRBEP は単調増加しており (図 1: 31.6→41.3，図 2: 35.6→46.2)，$\alpha = 0.2$〜1.0 では，総じて一定の範囲を浮動している (図 1: 34.7〜42.6，図 2: 40.3〜47.0)．また，図 2 は図 1 の変移グラフを縦軸正方向に移動したような形になっており，各 α における差分の平均値は約 4.7 である．

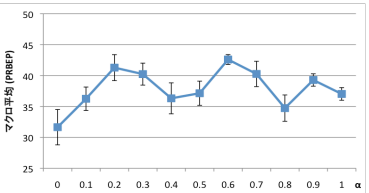

図 1: 一律に決定した最適パラメータを用いた際の平均 PRBEP

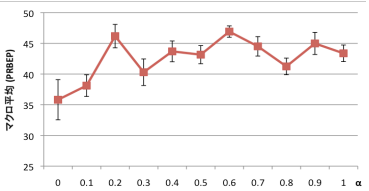

図 2: 各カテゴリ毎に最適パラメータを用いた際の平均 PRBEP

3.3 考察

図 1, 2 における全体的な変移パターンが類似していること，しかも，図 2 は図 1 のものよりおよそ一定幅上方にあること，これらのことは，グラフ構成に際して，一律に最適パラメータを決定するよりも，各カテゴリに対してそれぞれ最適パラメータを決定する方が優れていることを明示している．

また，図 1, 2 における $\alpha = 0$〜0.2 での PRBEP の増加は，取り入れた潜在情報の割合に応じて GBSSL 法の精度が向上することを示しているが，$\alpha = 0.2$〜1.0 では PRBEP は一定の範囲を浮動しており，両者に単純な相関を考えることは難しい．つまり，潜在情報を加味すればそれだけ精度向上に繋がるわけではない．データの内実に応じて，最大の精度向上をもたらす α の値が決まると考えられる．今後の課題としたい．

いずれにせよ，どの点も $\alpha = 0$ のときの点を上回っているので，教師データ選出に際しては，潜在情報を取り入れることが精度の向上に繋がることが分かる．

4 結論

GBSSL 法におけるグラフ構成に際しては，最適パラメータは一律に設定するよりも各カテゴリ毎に設定する方が優れている．また，教師データの選出に関連して本研究で提起した手法 (文書間の潜在情報を加味した教師データ選出による文書分類) は，精度向上に寄与し，有効である．

参考文献

[1] A. Subramanya and J. Bilmes, “Soft-Supervised Learning for Text Classification”, In *EMNLP*, 2008.

[2] 北島 理沙, 小林 一郎「潜在的意味を考慮したグラフに基づく複数文書要約」*Proceeding of ARG WI2*, 2012.

[3] D. M. Blei et al., “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[4] D. Zhou et al., “Learning with local and global consistency”, In *NIPS 16*, 2004.

[5] 江里口 瑛子, 小林 一郎「潜在情報を考慮したグラフに基づく半教師あり学習によるテキスト分類」情報処理学会第 75 回全国大会, 2013.

[6] 江里口 瑛子, 小林 一郎「潜在情報を考慮したグラフに基づく教師データの選出によるラベル伝搬法」言語処理学会第 19 回年次大会, 2013.