

# グラフに基づく半教師あり学習のための 潜在情報を考慮したグラフ構成

Graph Construction for Graph-based Semi-Supervised Learning using Latent Topics

江里口 瑛子      小林 一郎  
Akiko Eriguchi      Ichiro Kobayashi

お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻  
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

We aim to raise the accuracy of multi-class document categorization by means of graph-based semi-supervised learning (GBSSL). With this end in view, we propose three methods: the first one is a method to construct a similarity graph by employing both surface information and latent information to express similarity between nodes, the second is a method to use PageRank algorithm and select high-quality training data for GBSSL, and the third is to use the mutual  $k$ -nearest neighbor graph for GBSSL. We experimented on Reuters-21578 corpus and have confirmed that our proposed methods work well for raising the accuracy of multi-class document categorization.

## 1. 序論

機械学習手法には、教師あり学習、教師なし学習、半教師あり学習などがある。その中でも、グラフ構造に基づく半教師あり学習 (Graph-Based Semi-Supervised Learning: GBSSL) 法は、サポートベクターマシンの (SVM) などの学習法と比べてより有効な手法であることが知られている [Subramanya 08].

GBSSL 法の精度は、一方で、グラフ構成の仕方によって左右され、他方で、どのような教師データ (ラベルありデータ) を与えるかによっても左右される [Zhu 05, Zhu 03a]. 前者に関連して重要となるのは、グラフのノード間の関係性をどのように表現するかである。後者に関連して重要となるのは、情報量の大きい教師データをどのように選出するかである。この点に着目した分野が能動学習法の研究分野であり、GBSSL 法の精度を向上させるために、質の高い教師データの選出法がいくつか提案されている [Zhu 03a]. そのようなものの 1 つとして、我々は、文書分類タスクにおいて PageRank [Brin 98] 手法を用いた教師データの選出法を提案したことがある [江里口 13].

GBSSL 法で用いるグラフとして、各頂点をその  $k$ -近傍点集合と結びつけて構成した  $k$ -近傍グラフや、全ての頂点の次数を一定の値として作成した  $b$ -マッチンググラフ [Jebara 06] など様々なグラフが提案されている。しかし、いずれも難点がないわけではない。前者の  $k$ -近傍グラフには、次数の高い頂点 (ハブ) が作られやすいという傾向があり、このハブは GBSSL 法の分類精度を低下させることが知られている [Ozaki 11]. また、後者の  $b$ -マッチンググラフは精度向上に寄与するが、グラフ構成に必要な計算量がかなり大きく悩ましい点でもある。いずれも一長一短がある。これらの難点を解消する方法の一つに、相互  $k$ -近傍グラフを用いる方法がある。相互  $k$ -近傍グラフは、全ての頂点に対して次数制限が設けられ、ハブができにくくなっており、このグラフ構成を用いれば、 $b$ -マッチンググラフと同等の精度向上が得られ、しかも、グラフ構成に必要な計算量は少なく済む、という利点がある。Ozaki らは、語義曖昧性解消タスクにおいてこの手法の有効性を確認している [Ozaki 11].

本研究は、この方法が多クラス文書分類のタスクにおいて

も有効であることを示そうとするものである。その際、教師データの選出においては、文書の表層情報並びに潜在情報の両情報を用いる手法 [江里口 13] を用いることとし、上記方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、PRBEP (Precision-Recall Break Even Point) を算出し、相互  $k$ -近傍グラフを用いた手法の有効性を各カテゴリ毎に評価し、かつ、それら全体の精度の向上を検討する。この目的を果たすため、我々は、GBSSL 法で用いる分類用のグラフ構成について、 $k$ -近傍グラフと相互  $k$ -近傍グラフの 2 種類のグラフ構成を扱い、実験により、前者のグラフ構成を用いたときの結果をベースラインとして、後者のグラフ構成の手法が精度向上に資することを示す。

## 2. グラフに基づく文書分類

本研究で提示する、多クラス文書分類のタスクにおける GBSSL 法の詳細は、以下に述べる通りである。

### 2.1 グラフ構成

本研究のグラフ構成においては、テキストデータを対象に行う。したがって、各文書はグラフのノードとみなされる。そのノード (文書) 間の関係は類似度として表され、その類似度をグラフの辺の重みとするような重み付き無向グラフ  $G = (V, E)$  を構成する。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  $\mathbf{W}$  の形で表現することができ、 $w_{ij} \in \mathbf{W}$  はノード  $i$ 、ノード  $j$  間の類似度を表すとする。特に、GBSSL 法の場合には、その類似度はノード  $i$  の (相互)  $k$ -近傍点集合  $K(i)$  からなるものとし、 $w_{ij} = \text{sim}(x_i, x_j) \delta(j \in K(i))$  とする。ここで、 $\delta(z)$  は  $z$  が真ならば 1、偽ならば 0 とする。

### 2.2 ノード間の類似度

テキストデータにおける文書間の類似度を測る指標として、表層情報に基づく類似度と潜在情報に基づく類似度の二種類の類似度を採用する。文書の表層情報としては、文書に含まれる単語の出現頻度に着目した  $tfidf$  ベクトル [Salton 83] が多く用いられる。ここでは、表層情報に基づく類似度を、 $tfidf$  ベクトルのコサイン類似度 ( $\text{sim}_{\text{cos}}$ ) の値とする。また、文書の潜在情報としては、複数文書内に隠れトピックが存在することを仮定し、その隠れトピックに関して生起する単語の確率分布 (トピック分布) を用いる。ここでは、潜在情報に基づく類似度を、式 (2) によって得られる値 ( $\text{sim}_{JS}$ ) とし、トピック分布間の距離は Jensen-Shannon ダイバージェンス ( $D_{JS}$ ) を用いて求

連絡先: 江里口 瑛子, お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース 小林研究室, 〒112-8610 東京都文京区大塚 2-1-1, g0920506@is.ocha.ac.jp

める。トピック分布の推定には、Latent Dirichlet Allocation (LDA) 法 [Blei 03] を用いる。

本研究では、この従来の類似度 ( $sim_{cos}$ ) に新たに、文書の持つ潜在情報に基づいた類似度 ( $sim_{JS}$ ) を  $\alpha (0 \leq \alpha \leq 1)$  の割合で付加する。これら  $sim_{JS}$  と  $sim_{cos}$  を  $\alpha : (1-\alpha) (0 \leq \alpha \leq 1)$  の割合で合算した値を、ノード間 (すなわち、文書 S と文書 T 間) の類似度 ( $sim_{nodes}$ ) とする (式 (1))。P と Q は、それぞれ文書 S と文書 T に対するトピック分布を表す。

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

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

### 2.3 グラフを用いた質の高い教師データの選出

質の高い教師データの選出法として、複数文書自動要約タスクにおいて北島らが提案した TopicRank 法 [北島 12] を採用して行う。TopicRank 法とは、グラフ構造を用いた重要文抽出法の一つである。北島らは類似度グラフのノードを単文としたが、本研究では、類似度グラフのノードを単文から文書 (文の集合) に置き換えて用いる。辺の重みを文書間の潜在情報に基づく類似度として構成したグラフに対して、PageRank の概念を用いて式 (3) により各ノード (各文書) の重要度を算出し、各ノードの順位付けを行う。ここで、 $d$  は制動係数 (damping factor) である。式 (3) において、 $N$  を対象文書群の総文書数、 $adj[u]$  を文書  $u$  の隣接ノード集合とする。 $sim_{nodes}(u, v)$  は、式 (1) によって求めた文書  $u$  と文書  $v$  の類似度である。その上で、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 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。

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

上記手法により選出した教師データを用いてテストデータのラベルを予測する。

### 2.4 ラベル伝搬法

テストデータのラベルを予測する手法としては、ラベル伝搬法 [Zhou 04, Zhu 03b] を採用する。ラベル伝搬法は、カテゴリラベル未知のノード (テストデータ) について予測を行う手法である。予測値  $f$  は、以下の最適化問題の目的関数 (式 (4)) の解 (式 (6)) として求まる。 $W$  は類似度行列、 $n$  はノード数 (このうち、 $l$  は教師データ数)、 $\lambda (> 0)$  は両項のバランスをとる定数を表す。式 (4) は  $L$  を用いて、式 (5) と変形できる。 $L (\equiv D - W)$  はラプラシアン行列であり、 $D$  は  $W$  の各行 (又は列) の和を対角成分に持つ対角行列である。

$$J(f) = \sum_{i=1}^l (y^{(i)} - f^{(i)})^2 + \lambda \sum_{i < j} w^{(i,j)} (f^{(i)} - f^{(j)})^2 \quad (4)$$

$$= \|y - f\|_2^2 + \lambda f^T L f \quad (5)$$

$$f = (I + \lambda L)^{-1} y \quad (6)$$

### 2.5 相互 $k$ -近傍グラフ

相互  $k$ -近傍グラフとは、任意の 2 頂点間において互いに互いの  $k$ -近傍点集合に含まれているときのみ辺を作ること

得られるグラフである。しかしながら、この構成法ではグラフの全域性が欠如する可能性があるため、全域グラフ化を施したグラフを、本研究で用いる相互  $k$ -近傍グラフとして用いる。

## 3. 実験

### 3.1 実験仕様

テキスト分類問題の対象データには、Reuters-21578 (Reuters)\*<sup>1</sup> を用いる。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。本実験では “ModApte” 分割にしたがって、本文とタイトルのみからなる記事データを抽出し、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、同じデータセットを用いて GBSSL 手法でマルチラベル文書分類を行っている Subramanya ら [Subramanya 08] の実験仕様に合わせて、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。Reuters の記事データはマルチラベルを有するため、ここでは各カテゴリ毎に one-versus-rest 法を適用した二値分類を行い、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。

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  $l = 20$  個を加えたものを 11 セット用意する。データセットに含まれるデータ総数は  $n = 3319$  個である。教師データとして加えるカテゴリは、上記 10 種のカテゴリにそれら以外のカテゴリ (**others**) を加えた全 11 種とする。データセットに加える教師データ  $l$  個のカテゴリは 11 種のカテゴリからランダムに選択するが、全 11 種のカテゴリの教師データが少なくとも 1 個ずつ含まれるように選択する。

TopicRank 法を用いる際の LDA 法における潜在トピックの推定方法には、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。また、TopicRank 法で用いるグラフは、ノード数  $|V| = (\text{カテゴリ毎の教師データの総数})$ 、辺数  $E = |V \times V|$  の完全グラフとする。パラメータ  $\alpha$  は、0.0 から 1.0 まで 0.1 刻み毎の値を与え、制動係数  $d$  は Brin らの結果を参考に 0.85 とする [Brin 98]。カテゴリ毎に各文書の TopicRank スコアを算出し、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、推定を行う必要がない。このため、類似度が一意的に決まるのでスコアは 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の推定を行うため、類似度が一意的に決まらない。このため、5 回平均の値をスコアとする。

ラベル伝搬法で用いた類似度グラフのノード数は  $|V| = n (= 3319)$  であり、ノード間の類似度は、パラメータ  $\alpha = 0$  とし、表層情報のみからなるものとする。 $k$ -近傍グラフ、相互  $k$ -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  $\lambda$  は  $\{1, 0.1, 0.01, 1e-4, 1e-8\}$  の範囲を動かす。最初のデータセットによって、各カテゴリに対する最適パラメータ ( $k, \lambda$ ) の組を決定した後、それらのパラメータの値を用いて、残り 10 セットに対して文書分類を行い、各カテゴリ毎に PRBEP を求め、各試行毎の各カテゴリに対する PRBEP の平均値を算出する。指標 PRBEP は、Precision (適合率) と Recall (再現率) が一致するときの値である。 $k$ -近傍グラフを用いた際の最適パラメータ ( $k, \lambda$ ) は [江里口 13] を採用する。

\*1 <http://www.daviddlewis.com/resources/testcollections/reuters21578/>

表 1: 相互  $k$ -近傍グラフを用いた際の, カテゴリ毎の最適パラメータ ( $k, \lambda$ )

カテゴリ \ $\alpha$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
earn	(500, 1)	(500, 1)	(500, 1)	(500, 1)	(500, 1)	(250, 1)	(250, 1)	(500, 1)	(250, 1)	(500, 1)	(250, 1)
acq	(500, 1e-8)	(3318, 1)	(500, 1e-8)	(2, 0.1)	(10, 0.1)	(2000, 0.1)	(3318, 0.1)	(2000, 0.1)	(3318, 0.1)	(2000, 0.1)	(2000, 0.1)
money-fx	(500, 1e-4)	(2, 1)	(50, 1e-8)	(10, 1)	(500, 1e-8)	(500, 1e-8)	(500, 1e-8)	(10, 1)	(500, 1e-8)	(10, 1)	(10, 1)
grain	(1000, 1)	(1000, 0.01)	(100, 1)	(50, 1)	(100, 1)	(50, 1)	(1000, 0.01)	(100, 1)	(100, 1)	(100, 1)	(50, 1)
crude	(50, 1e-8)	(250, 1e-8)	(100, 0.1)	(100, 1e-8)	(100, 1e-4)	(100, 1e-8)	(500, 1e-8)	(500, 1e-8)	(100, 1e-8)	(250, 0.01)	(100, 1e-8)
trade	(2, 1)	(10, 0.1)	(10, 1)	(100, 0.1)	(50, 1e-8)	(10, 0.1)	(50, 1e-8)	(10, 0.1)	(50, 1e-4)	(10, 0.01)	(10, 0.01)
interest	(100, 0.01)	(100, 0.01)	(250, 1e-8)	(50, 0.01)	(500, 1e-8)	(500, 1e-8)	(500, 1e-8)	(250, 1e-8)	(500, 1e-8)	(250, 1e-8)	(500, 1e-8)
ship	(100, 1e-8)	(50, 1)	(250, 0.01)	(250, 0.1)	(100, 0.01)	(250, 0.01)	(100, 0.1)	(100, 0.1)	(100, 1)	(500, 0.01)	(250, 0.01)
wheat	(1000, 1e-8)	(500, 0.01)	(250, 1e-8)	(100, 0.1)	(1000, 1e-4)	(1000, 0.01)	(500, 1e-8)	(500, 1e-8)	(500, 1e-8)	(500, 1e-8)	(500, 1e-8)
corn	(10, 1e-8)	(500, 1e-8)	(500, 1e-8)	(250, 1e-8)	(250, 1e-8)	(500, 1e-8)	(250, 1e-8)	(500, 1e-4)	(500, 1e-8)	(250, 1e-8)	(500, 1e-8)

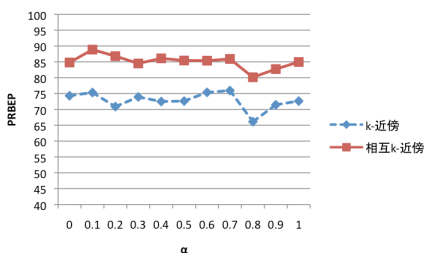


図 1: earn の平均 PRBEP

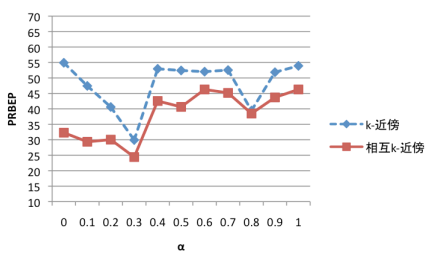


図 2: acq の平均 PRBEP

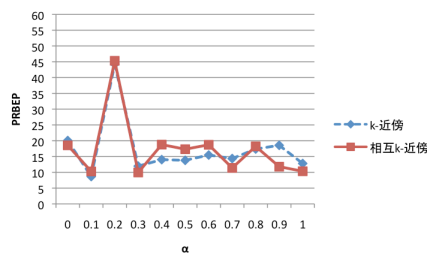


図 3: money-fx の平均 PRBEP

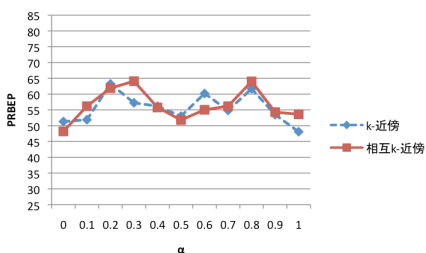


図 4: grain の平均 PRBEP

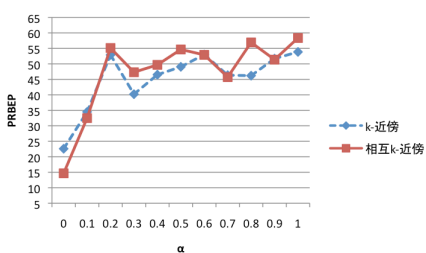


図 5: crude の平均 PRBEP

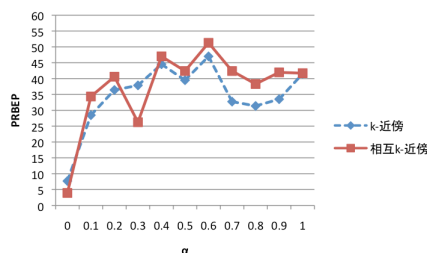


図 6: trade の平均 PRBEP

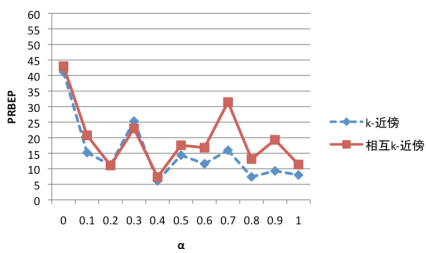


図 7: interest の平均 PRBEP

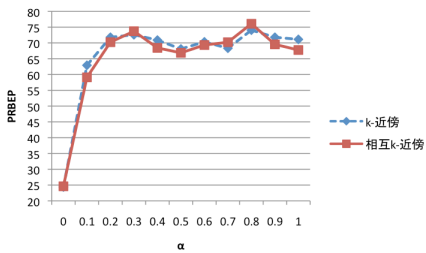


図 8: ship の平均 PRBEP

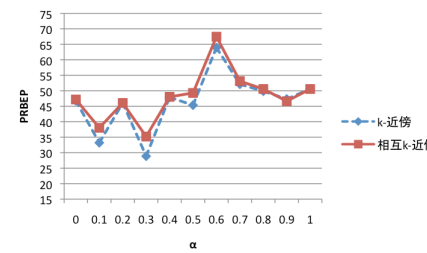


図 9: wheat の平均 PRBEP

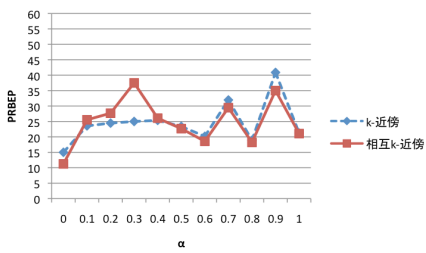


図 10: corn の平均 PRBEP

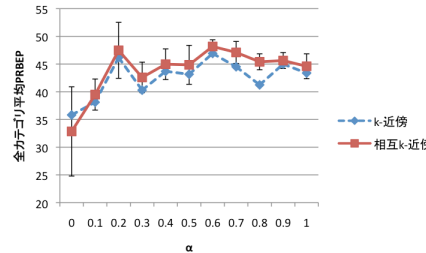


図 11: 全カテゴリの平均 PRBEP

表 2: ベースラインとの差の平均値

カテゴリ	平均値
earn	+12.23
acq	-9.89
money-fx	-0.06
grain	+0.90
crude	+2.03
trade	+2.69
interest	+4.46
ship	-0.95
wheat	+1.83
corn	+0.28

### 3.2 実験結果

表1は、相互  $k$ -近傍グラフを用いた際の、 $[0, 1]$  における0.1刻み毎の各  $\alpha$  の値に対してカテゴリ毎に決定した最適パラメータ  $(k, \lambda)$  を表す。図1~10は、各カテゴリに対し、これらの最適パラメータを用いて行った実験結果を折れ線で結んだものである。図中の青色の点線は  $k$  近傍グラフを用いた結果 (ベースライン) であり、赤色の実線は相互  $k$ -近傍グラフを用いた結果である。横軸は、表層情報に対する潜在情報の混合割合  $\alpha$  の値を表し、縦軸は PRBEP の値 (精度) を表す。図1~10は、各  $\alpha$  の値に対する10回の試行の各カテゴリ PRBEP の平均値を示している。加えて、全  $\alpha$  におけるベースラインとの差の平均値を表2に示した。図11は、全カテゴリの PRBEP を各  $\alpha$  毎に合算して求め、その平均値 (マクロ平均値) を示している。図中のエラーバーは、標準偏差の大きさを表す。

図における  $\alpha = 0, 1$  の場合は、それぞれ、表層情報のみ、潜在情報のみを用いた場合の結果を示す。それ以外 ( $\alpha \neq 0$  または  $1$ ) は、潜在情報と表層情報を一定の割合 ( $\alpha : (1 - \alpha)$ ) で混合した場合であり、両情報を用いた結果を示している。

各カテゴリ毎の図1~10に関連して、次のことが言える。図1においては、相互  $k$ -近傍グラフを用いた時の精度 (PRBEP の値) が全ての  $\alpha$  においてベースラインの上方にあり、他方、図2においては全て下方にある。図3~10では、 $\alpha$  の値によってベースラインの上方または下方になっている。表2によると、ベースラインとの差の平均値が正の値をとっているものが、全カテゴリ10個のうち7個あり、総合して見ると、全体として精度向上を期待することができる。

図11からは、次のことが分かる。ベースラインのマクロ平均値の最大値は46.9 ( $\alpha = 0.6$  の時) であり、最小値は35.8 ( $\alpha = 0$ ) である。 $\alpha = 1$  の時の値は43.4である。相互  $k$ -近傍グラフを用いた際の、マクロ平均値の極大値は47.5, 45.0, 48.2, 45.6 (それぞれ  $\alpha = 0.2, 0.4, 0.6, 0.9$ ) であり、最大値は48.2 ( $\alpha = 0.6$ ) である。最小値は32.8 ( $\alpha = 0$ ) である。 $\alpha = 1$  の時の値は44.6である。また、各  $\alpha$  毎のマクロ平均値は、ベースラインよりも平均して1.4%高く、特に  $\alpha = 0.8$  では、最大で4.2%も高くなっている。さらに、精度 (PRBEP) が最大となる  $\alpha = 0.6$  においてベースラインとの差異の有無を有意水準5%で  $t$  検定を行うと、有意な差で精度向上のあることが知れた。

### 4. 考察

相互  $k$ -近傍グラフを用いる方法の精度は、図1(earn)を除いて、 $\alpha$  の値によってはベースラインを下回っているものもあり、各カテゴリの諸結果からだけでは、精度の向上を単純に論定することはできない。しかしながら、各カテゴリ毎の PRBEP のマクロ平均を比較してみると、相互  $k$ -近傍グラフを用いる方法の精度 (PRBEP) は、 $\alpha \neq 0$  ときの全てにおいてベースラインよりも高く、平均して1.4%上回っている。尤も、 $\alpha = 0 \sim 0.5, 1.0$  では誤差範囲が大きく、優劣を論じることができないが、 $\alpha = 0.6 \sim 0.9$  では明らかに有意の差で精度向上が見られる。したがって、 $\alpha = 0.6 \sim 0.9$  の範囲の限りで、相互  $k$ -近傍グラフを用いる方法は精度が高くなると論定できよう。ちなみに、 $\alpha = 0.6$  においては、相互  $k$ -近傍グラフを用いる方法並びにベースライン共々、最大値をとっている点は注目に値する。いずれのグラフ構成の方法を用いたとしても、教師データを選出する手法としては、表層情報と潜在情報の両情報を用いる方が、精度が高くなることを示しているからである。

図11における  $\alpha = 0$  の時に、我々の用いた方法の精度がベースラインよりも低くなっていることに関しては、次の理由が考えられる。 $\alpha = 0$  の時は、表層情報のみに基づいて教師

データを選出した場合に相当し、このとき選出されやすいのは、他のデータと高い類似度で多く結ばれている (頂点次数の高い) データである。相互  $k$ -近傍グラフを用いる際には、頂点次数にある一定の上限があり、対して、 $k$ -近傍グラフを用いる際には、そのような一定の制限がなく、選出された教師データは、グラフにおける頂点次数の高いハブとなりやすく、その分だけ精度向上への寄与が見込まれる。相互  $k$ -近傍グラフでは、この寄与分が減殺されるため、ベースラインを下回る。

### 5. 結論

多クラス文書分類タスクにおいては、教師データを選出する際、文書の表層情報と潜在情報のある一定の混合割合で用いるのが良く、このとき、GBSSL法のグラフ構成には相互  $k$ -近傍グラフを用いる方法が精度向上に資する。

今後の課題としては、他のデータセットを用いて同様の結果が得られるかを検証することであり、最適パラメータ  $(k, \lambda)$  における決定の仕方を改善し、精度向上に努めることである。

### 参考文献

- [Blei 03] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research* (2003)
- [Brin 98] Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine., *Computer Networks and ISDN Systems*, pp. 107-117 (1998)
- [Jebara 06] Jebara, T., Shchogolev, V.: B-Matching for spectral clustering, in *Proc. of ECML* (2006)
- [Ozaki 11] Ozaki, K., Shimbo, M., Komachi, M., Matsumoto, Y.: Using the Mutual  $k$ -Nearest Neighbor Graphs for Semi-supervised Classification of Natural Language Data, in *Proc. of ACL* (2011)
- [Salton 83] Salton, G., McGill, J.: Introduction to Modern Information Retrieval, McGraw-Hill (1983)
- [Subramanya 08] Subramanya, A., Bilmes, J.: Soft-Supervised Learning for Text Classification, in *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp.1090-1099 (2008)
- [Zhou 04] Zhou, D., Bousquet, O., Lal, T. N., Weston J., Schölkopf B.: Learning with Local and Global Consistency, in *NIPS 16* (2004)
- [Zhu 02] Zhu, X., Ghahramani, Z.: Learning from Labeled and Unlabeled Data with Label Propagation, Technical report, Carnegie Mellon University (2002)
- [Zhu 03a] Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, in *Proc. of the International Conference on Machine Learning (ICML)* (2003)
- [Zhu 03b] Zhu, X., Ghahramani, Z., Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions, in *ICML* (2003)
- [Zhu 05] Zhu, X.: Semi-Supervised Learning with Graphs, PhD thesis, Carnegie Mellon University (2005)
- [江里口 13] 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)
- [北島 12] 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, in *Proc. of ARG WI2* (2012)