自由エネルギー距離によるハブの抑制

小林 雄太　　新保 仁　　松本 裕治
奈良先端技術大学院大学 情報科学研究科
{kobayashi.yuta.kp1, shimbo, matsu}@is.nasit.jp

1 はじめに

1.1 背景

近傍法 (nearest neighbor method) は機械学習をはじめ自然言語処理などの種々の分野で使用される代表的な手法の一つである. 例えば，自然言語処理では文書分類や語義曖昧性解消，品詞タグ付けなどで用いられる. 近傍法を用いた分類問題では，クラスラベルが未知のテスト事例のオブジェクトに対してその近傍を計算し，近傍に含まれるオブジェクトのクラスラベルの投票によって分類を行う.

近年，近傍法の問題の 1 つとして，高次元データにおけるハブの出現が注目されている [4][5]．ハブとは多数のオブジェクトと類似するオブジェクトのことであり，近傍法におけるハブの出現は多数の近傍リストに同一オブジェクトが含まれることを意味する．分類問題において，ハブのラベルはテスト事例のオブジェクトに付与されやすくなるため，結果として近傍法を用いた分類精度の低下を引き起こす原因となる. ハブによる近傍法の精度低下は以前から報告されていたが, Radovanović らはハブの出現が次元の呪いの一側面であることを理論的に示した [4].

また，自然言語処理分野などの高次元データから構築されたグラフにおいてもハブが出現し，半教師あり分類の精度を低下させることが知られている [3]．したがって，これらの手法は適切な距離尺度，すなわちハブの出現を抑える距離尺度を使用する必要がある．しかしながら，Luxburg らはグラフ節点間の距離尺度の一つである通勤時間距離 (commute time distance) を大規模グラフに対して用いた場合，グラフの大域的な構造を反映しない尺度になってしまうことを理論的に示した [6]．その際，通勤時間距離は 2 節点の次数の逆数の和に近似できるため，どの節点の近傍リストも同じになると述べている．これはすなわち，次数の大きい節点がハブとなることに他ならない．一方で，通勤時間距離と並び最も標準的なグラフ距離尺度の最短経路距離や，最短経路距離と通勤時間距離を一般化した自由エネルギー距離 (free energy distance) [2] において，ハブが出現するかどうかは明らかにされていない.

1.2 研究目的

本稿ではグラフ距離尺度である最短経路距離と自由エネルギー距離を使用し，ハブの抑制効果を典型的な近傍探索の一例である文書分類タスクにより検証する．このタスクにおいて文書は一般的に bag-of-words の高次元ベクトルによって表現され，文書間の類似度をこのベクトルを用いて計算する．そして，類似度から 10 近傍グラフを構築しグラフ距離尺度を適用することで文書間類似度を得る．その後，テスト文書と最も類似度の高い k 個の訓練文書のラベルによる重み付き投票によりテスト文書の未知ラベルを推定する．先行研究において，文書分類データにハブが出現していることが報告されており [3][5]，本稿では文書分類に自由エネルギー距離を適用することで，ハブの出現を抑え近傍分類の精度を向上させられるか検証した.

1.3 貢献

本稿の貢献を以下に示す.

- グラフ距離尺度である最短経路距離や自由エネルギー距離とハブ度の関係を調査した．先行研究では通勤時間距離がハブの出現を引き起こすことが報告されていたが [6]，最短経路距離やそれらの自由エネルギー距離がハブを抑制可能かは明らかにされていなかった．文書分類実験において，自由エネルギー距離によりハブの出現が抑制され，分類精度が向上することを確認した．元の文書間類似度，通勤時間距離や最短経路距離を用いた場合のいずれと比較してもハブが抑制され，特に訓練事例が少ない場合に精度が大きく向上し，効果的な手法であると確認した.
- ハブ度の尺度である N_{10} 歪度により最適な自由エネルギー距離のパラメータ β を設定できる可能性を示した．ハブ度はオブジェクトのクラスラベルなしで計算可能というメリットがある.

2 グラフの距離尺度

2.1 最短経路距離

グラフ距離尺度として最も一般的なものは最短経路距離である．本稿でのグラフは，文書間類似度に基づいて生成された重み付き無向グラフを取り扱うため，グラフの 2 節点間の全経路のうち，総コストが最小となる経路を最短経路距離とする．計算アルゴリズムには Floyd-Warshall 法を使用した [1]．最短経路距離は，節点間の経路数といったグラフ全体の構造を考慮しないという欠点を持つ．

2.2 通勤時間距離

通勤時間距離は最短経路距離と並んで一般的なグラフ距離尺度である．通勤時間距離は節点 i から出発したランダムウォーカーが，節点 j に初めて到着し，そこから初めて i に戻るのに要する時間の期待値である．N 節点からなるグラフ G について，各節点間の重み $a_{i,j}$ を (i,j) 要素に持つ隣接行列 $\mathbf{A}$ が与えられたとき，$\mathbf{L}=\mathbf{D}-\mathbf{A}$ で定義される行列 $\mathbf{L}$ を，G のラプラシアン (Laplacian) と呼ぶ．なお，$\mathbf{D}$ は対角要素 $\mathbf{D}(i,i)$ が $\mathbf{A}$ の i 行の要素和となる対角行列である．最終的に通勤時間距離 CT は次式で表される．

$$\mathrm{CT}(i,j) = v_g\,(\mathbf{e}_i-\mathbf{e}_j)^{\mathrm{T}}\mathbf{L}^{\dagger}(\mathbf{e}_i+\mathbf{e}_j)$$

ここで，$\mathbf{L}^{\dagger}$ は $\mathbf{L}$ の疑似逆行列，v_g はグラフ G の volume（合計節点次数）であり，$\mathbf{e}_i$ は第 i 要素のみ 1，それ以外が 0 のベクトルである．通勤時間距離は，節点間の経路数といったグラフ全体の構造を考慮する点は長所だが，大規模なグラフでは距離尺度として適切に機能しない短所を持つ [6].

2.3 自由エネルギー距離

自由エネルギー距離は Kivimäki ら [2] によって提案された，最短経路距離と通勤時間距離をパラメータ化により一般化した，グラフの 2 節点間の距離尺度である．自由エネルギーの概要について簡単に述べる．

まず，隣接行列とは別に，コスト行列 $\mathbf{C}$ は (i,j) 要素に i,j 節点間の辺のコスト $c_{i,j}$ を持つ．$c_{i,j}$ は $a_{i,j}$ と独立に定めることができるが，ここでは $a_{i,j}$ の逆数 $c_{i,j}=1/a_{i,j}$ を用いる．また，各経路 $\wp$ のコスト $\widetilde{c}(\wp)$ を，その経路に含まれる辺のコストの総和として次式で定義する．

$$\widetilde{c}(\wp) = \sum_{(i,j)\in\wp} c_{ij}$$

グラフ上の 2 節点 s,t 間の全経路（ただし，途中 t は訪問しない経路に限る）を $\mathcal{P}_{st}$ とする．一般に $\mathcal{P}_{st}$ は可算無限集合である. 自由エネルギー距離では，$\mathcal{P}_{st}$ 上の確率分布 $\widetilde{P}_{st}(\wp)$ によって各経路 $\wp\in\mathcal{P}_{st}$ に確率を割り振る事を考える．ここで，$\wp$ は節点 s から節点 t までの経路を表し，$\wp=(i_1=s,\ldots,i_T=t)\in\mathcal{P}_{st}$ である．ただし，$i_j\neq t\ \forall j=1,\ldots,T-1$.

$\widetilde{P}_{st}(\wp)$ の一例に，辺の重み $a_{i,j}$ を反映した「自然な」ランダムウォークによる確率分布 $\widetilde{P}^{\mathrm{ref}}_{st}(\wp)$ がある．ランダムウォーカーが節点 i から節点 j に遷移する確率を表す遷移行列 $\mathbf{P}^{\mathrm{ref}}$ の (i,j) 要素 $p^{\mathrm{ref}}_{i,j}$ は次式で計算される．

$$p^{\mathrm{ref}}_{i,j} = \frac{a_{i,j}}{\sum_k a_{i,k}}$$

したがって，自然なランダムウォークによる経路 $\wp$ の遷移確率分布 $\widetilde{P}^{\mathrm{ref}}_{st}(\wp)$ は $\widetilde{P}^{\mathrm{ref}}_{st}(\wp)=p^{\mathrm{ref}}_{i_1i_2}\cdots p^{\mathrm{ref}}_{i_{T-1}i_T}$ となる．

さて，任意の分布 $\widetilde{P}_{st}(\wp)$ のもとでの経路コストの期待値 $\bar{c}(\widetilde{P}_{st})$ は次式で与えられる．

$$\bar{c}(\widetilde{P}_{st}) = \sum_{\wp\in\mathcal{P}_{st}} \widetilde{P}_{st}(\wp)\widetilde{c}(\wp)$$

一方，$\mathcal{P}_{st}$ 上の確率分布 $\widetilde{P}_{st}(\wp)$ と自然ランダムウォークによる遷移確率分布 $\widetilde{P}^{\mathrm{ref}}_{st}(\wp)$ の相対エントロピー $J(\widetilde{P}_{st}(\wp)\|\widetilde{P}^{\mathrm{ref}}_{st}(\wp))$ は，逆温度パラメータ β を用いて重み付けすると次式で表される．

$$\frac{1}{\beta}J(\widetilde{P}_{st}(\wp)\|\widetilde{P}^{\mathrm{ref}}_{st}(\wp)) = \frac{1}{\beta}\sum_{\wp\in\mathcal{P}_{st}}\widetilde{P}_{st}(\wp)\log(\widetilde{P}_{st}(\wp)/\widetilde{P}^{\mathrm{ref}}_{st}(\wp))$$

コストの期待値 $\bar{c}(\widetilde{P}_{st})$ と重み付けした相対エントロピー $\frac{1}{\beta}J(\widetilde{P}_{st}(\wp)\|\widetilde{P}^{\mathrm{ref}}_{st}(\wp))$ の和の最小化問題を，$\sum_{\wp\in\mathcal{P}_{st}}\widetilde{P}_{st}(\wp)=1$ の制約のもとで解くと次式を得る．

$$\widetilde{P}^{FE}_{st}(\wp) = \frac{\widetilde{P}^{\mathrm{ref}}_{st}(\wp)\exp\left(-\beta\widetilde{c}(\wp)\right)}{\sum_{\wp\in\mathcal{P}_{st}}\widetilde{P}^{\mathrm{ref}}_{st}(\wp)\exp\left(-\beta\widetilde{c}(\wp)\right)}$$

自由エネルギー距離はこの確率分布 $\widetilde{P}^{FE}_{st}$ と $\widetilde{P}^{FE}_{ts}$ により

$$\Delta^{FE}_{st} = \frac{\phi(\widetilde{P}^{FE}_{st}(\wp))+\phi(\widetilde{P}^{FE}_{ts}(\wp))}{2}$$

と定義される．ただし，

$$\phi(\widetilde{P}^{FE}_{st}) = -\frac{1}{\beta}\log\Big(\sum_{\wp\in\mathcal{P}_{st}}\widetilde{P}^{\mathrm{ref}}_{st}(\wp)\exp(-\beta\widetilde{c}(\wp))\Big)$$

詳しい導出は文献を参照されたし [2].

上式には無限の経路に対する加算が含まれているが，Algorithm 1 によって効率的に計算することが可能である．逆温度パラメータ β は経路コストをどの程度考慮するかを決定する．自由エネルギー距離は，$\beta\to 0^{+}$ において通勤時間距離の 1/2 倍と一致し，$\beta\to\infty$ において往復最短経路距離の 1/2 倍と一致する．

3 実験

グラフ距離尺度を使った文書分類タスクにより，分類精度とハブ度の関係を調べた．

Algorithm 1 グラフ G の全節点間の自由エネルギー距離の計算

Require:
– A graph G containing n nodes.
– n×n adjacency matrix $\mathbf{A}$ associated to $G = (V, E)$, representing affinities.
– $n \times n$ cost matrix $\mathbf{C}$ associated to $G = (V, E)$.
– Inverse temperature parameter β.

$\mathbf{D} \leftarrow \mathbf{Diag}(\mathbf{Ae})$ {row-normalization matrix}
$\mathbf{P}^{\mathrm{ref}} \leftarrow \mathbf{D}^{-1}\mathbf{A}$ {reference transition probabilities matrix}
$\mathbf{W} \leftarrow \mathbf{P}^{\mathrm{ref}} \circ \exp[-\beta\mathbf{C}]$ {elemen-twise exponential and multiplication}
$\mathbf{Z} \leftarrow (\mathbf{I} - \mathbf{W})^{-1}$ {fundamental matrix}
$\mathbf{Z}_h \leftarrow \mathbf{Z}\,\mathbf{Diag}(\mathbf{Z})^{-1}$ {fundamental matrix of hitting paths}
$\phi \leftarrow -\frac{1}{\beta}\log \mathbf{Z}_h$ {dissimilarity matrix by symmetrization}
$\mathbf{\Delta}^{FE} \leftarrow (\phi + \phi^{\mathrm{T}})/2$
return $\mathbf{\Delta}^{FE}$

3.1 データセット

文書分類実験のデータセットとして，20Newsgroups を用いた．20Newsgroups は 20 カテゴリの Usenet ニュースグループに投稿された 2 万弱の文書からなり，文書分類において標準的なデータセットである．本稿ではステミングとストップワード処理を行い，重複した文書を除いた公開データセット [*1] を使用した．データセットの全文書数は 18846 文書，bag-of-words 特徴量は 26214 次元，文書クラスは 20 クラスである．また，データセットには訓練データとして，全文書の 5, 10, 20, 30, 40, 50 %をランダムに選んだものが各 10 セットずつ用意されている．これらを訓練文書とし，訓練文書以外のデータをテスト文書として使用し分類精度を測定する．

3.2 素性から文書間距離への計算

文書間の類似度は，一般的な bag-of-words 特徴量のコサイン類似度により計算した．この類似度をもとに 10 近傍グラフを構築し，隣接行列を生成した．10 近傍グラフは全節点間に経路が存在する連結グラフであった．i, j 節点間の辺のコスト $c_{i,j}$ の計算は，慣例的に類似度 $a_{i,j}$ を用いて $c_{i,j} = 1/a_{i,j}$ とした．

3.3 比較手法

分類精度とハブ度の評価を以下の手法により行った．

- bag-of-words コサイン類似度 (Baseline)
- 10 近傍グラフ構築後，通勤時間距離を適用
- 10 近傍グラフ構築後，最短経路距離を適用
- 10 近傍グラフ構築後，自由エネルギー距離を適用

自由エネルギー距離のパラメータ β は
$\beta \in \{10^{-4}, 10^{-3.5}, \cdots, 10^{0}, \cdots, 10^{2}\}$ の 13 段階とした．

[*1] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

表 1: 距離尺度と分類精度 (%) の関係　(CT:通勤時間距離, FE：自由エネルギー距離, SP:最短経路距離)

訓練文書比率 (%)	CT	FE	SP	Baseline
5	7.8 (−50.4)	**66.5 (+8.3)**	63.5 (+5.3)	58.2
10	9.6 (−55.2)	**69.8 (+5.0)**	68.0 (+3.2)	64.8
20	11.0 (−60.1)	**74.5 (+3.4)**	73.0 (+1.9)	71.1
30	13.2 (−61.6)	**77.3 (+2.5)**	76.1 (+1.3)	74.8
40	14.5 (−62.4)	**79.1 (+2.2)**	78.3 (+1.4)	76.9
50	14.0 (−64.6)	**80.3 (+1.7)**	79.7 (+1.1)	78.6

3.4 評価

3.4.1 文書分類の評価尺度

文書分類タスクでは，距離尺度の適用後に重み付き k 近傍分類によるクラスラベル推定精度によって評価を行う．各テスト文書の k 近傍に含まれる訓練文書のクラスラベルを用いて，重み付き投票を行いクラスラベルを推定することで分類精度を計算した．投票の重みは文書間距離の逆数の 2 乗とした．k 近傍分類のパラメータ k は $k = \{1, \cdots, 10\}$ の 10 段階とした．

3.4.2 ハブ度の評価尺度

距離尺度の適用後に N_{10} 歪度を用いてハブ度の評価を行う．N_{10} 分布は各文書が，他文書の近傍上位 10 位以内に何回出現したかを表す分布であり，その歪度 N_{10} 歪度は手法がハブの影響をどれだけ受けているかを示す [4][5]．N_{10} 分布の歪度は次式で計算される．

$$(N_{10}\text{歪度}) = \frac{\sum_{i=1}^{n}(x_i - \mu)^3/n}{\sigma^3}$$

$$x_i = \sum_{i \neq j} 1/(d_{10}(i, j))^2$$

ここで，n は全文書数，μ は投票された重みの平均，σ は投票された重みの標準偏差である．一般に歪度は $x_i = 1$ であるが，本稿では重み付き k 近傍分類を行うため x_i の値を重み付きで考える．すなわち x_i は i 番目の文書が他文書の近傍上位 10 位以内に出現した場合の，各文書との距離 $d_{10}(i, j)$ の逆数の 2 乗の総和とした．N_{10} 歪度が小さいほど，様々な節点の近傍に出現するハブが少なく，ハブ度が低いことを意味する．

4 実験結果と考察

4.1 訓練文書数と k 近傍分類の精度

表 1 に各訓練文書比率における，距離尺度と分類精度の関係を示す．各距離尺度の分類精度の値は，k を 1 から 10 まで変化させた際の最高分類精度である．太字の値は各訓練文書比率における最高精度，括弧中の値は Baseline からの改善度を意味する．全ての訓練文

表 2: 自由エネルギー距離のパラメータ β と 10 近傍分類精度 (%) の関係　(CT:通勤時間距離, SP:最短経路距離)

訓練文書比率 (%)	CT	$\log_{10}\beta$													SP	Baseline
		-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2		
5	6.8	20.4	39.9	59.2	65.5	**66.5**	66.4	66.3	66.0	65.3	64.5	64.0	63.7	62.6	63.5	58.2
10	7.7	27.0	49.8	65.4	69.0	69.3	69.5	69.7	**69.8**	69.4	68.8	68.4	68.1	67.9	68.0	64.8
20	10.8	40.2	62.4	70.8	72.3	72.6	73.1	73.8	**74.5**	74.5	73.9	73.4	73.1	73.0	73.0	71.1
30	12.4	49.6	69.5	73.5	74.2	74.5	75.1	76.0	76.9	**77.3**	77.0	76.5	76.3	76.2	76.1	74.8
40	14.5	54.0	72.0	74.8	75.3	75.7	76.3	77.4	78.5	**79.1**	78.9	78.6	78.4	78.3	78.3	76.9
50	13.6	54.9	73.7	75.9	76.5	76.8	77.5	78.6	79.8	**80.4**	80.3	80.0	79.8	79.7	79.7	78.5

表 3: 自由エネルギー距離のパラメータ β と N_{10} 歪度の関係　(CT:通勤時間距離, SP:最短経路距離)

	CT	$\log_{10}\beta$													SP	Baseline
		-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2		
N_{10} 歪度	44.8	53.8	44.4	33.2	31.3	21.9	6.52	1.71	**0.886**	1.09	2.36	3.28	3.70	3.84	3.91	3.85

書比率において，自由エネルギー距離の分類精度が最も高く，特に訓練データが少ない状況下で自由エネルギー距離が効果的であることが確認できる．また，最短経路距離と通勤時間距離の精度の差から，最短経路距離が通勤時間距離に比べ距離尺度として適切に機能することを確認した．この結果は，先行研究の通勤時間距離がグラフの大域的な構造を反映しないという主張と合致する [6]．自由エネルギー距離が最高精度となった際，全ての訓練文書比率で $k = 10$ であったため，以降は $k = 10$ での分類精度について議論する．

4.2　ハブ度と k 近傍分類の精度

表 2 に各訓練文書比率における，自由エネルギー距離のパラメータ β と $k = 10$ での分類精度の変化を示す．太字の値は各訓練文書比率における最高分類精度を意味する．また，表 3 にパラメータ β による N_{10} 歪度の変化を示す．太字の値は N_{10} 歪度の最小値であり，最もハブが抑えられていることを意味する．

表 2 及び表 3 より，N_{10} 歪度が最小値を取る $\log_{10}\beta = -0.5$ 付近で k 近傍分類の精度が特に向上しており，ハブ度と k 近傍分類の精度が深く関係することを確認した．また，通勤時間距離と最短経路距離においても，ハブ度が分類精度と関係するとわかった．したがって，ハブ度の尺度である N_{10} 歪度により，自由エネルギー距離の最適なパラメータ β を設定できる可能性がある．

通勤時間距離が大規模グラフにおいて距離尺度として適切に機能しないという Luxburg らの先行研究から推測すると [6]，自由エネルギー距離が距離尺度として通勤時間距離から最短経路距離に近づくに従い，ハブ度が単調減少し分類精度が単調増加すると考えられる．しかしながら，実際には 2 つの距離尺度の中間で分類精度のピークを持つことを確認した．

5　まとめ

文書データセットから生成した大規模グラフに対してグラフ距離尺度を適用し，ハブの出現を抑え近傍分類の精度を向上させられるか検証した．文書分類実験の結果，自由エネルギー距離は適切なパラメータ設定によりハブを抑制し，分類精度の向上を確認した．また，N_{10} 歪度によるパラメータ設定の可能性を示した．

本論文は，最短経路距離が通勤時間距離と比べハブ抑制の観点で優れている事，2 つの距離尺度の中間をとる自由エネルギー距離が最もハブを抑制し，近傍探索の精度を向上させることを初めて示した．今後は 2 つの距離尺度の中間でハブが抑制される理論的な解析や，他のグラフ距離尺度についても検証したい．

参考文献

[1] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, Vol. 5, No. 6, p. 345, 1962.

[2] Ilkka Kivimäki, Masashi Shimbo, and Marco Saerens. Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A: Statistical Mechanics and its Applications*, Vol. 393, pp. 600–616, 2014.

[3] Kohei Ozaki, Masashi Shimbo, Mamoru Komachi, and Yuji Matsumoto. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. In *Proc. CoNLL*, pp. 154–162, 2011.

[4] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *The Journal of Machine Learning Research*, Vol. 11, pp. 2487–2531, 2010.

[5] Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Yuji Matsumoto, Marco Saerens, et al. Investigating the effectiveness of laplacian-based kernels in hub reduction. In *Proc. AAAI*, pp. 1112–1118, 2012.

[6] Ulrike von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, Vol. 15, pp. 1751–1798, 2014.