Vol.2010-MPS-77 No.19
2010/3/5

制約を反映するグラフ表現に基づく射影による半教師ありクラスタリング

吉田哲也[†1]　岡谷一宏[†1]

本稿では，must-link と cannot-link と呼ばれる制約が与えられる場合に対して，制約を反映するグラフ表現に基づく射影を用いて半教師ありクラスタリングを行う手法を提案する．提案手法ではデータ全体を類似度に基づいてグラフ構造として表現し，グラフ理論における縮約とグラフラプラシアンによる射影を用いてそれぞれの制約を反映した射影表現を構築し，構築した射影表現に対してクラスタリングを行う．提案手法を高次元スパースな表現を持つ実データに対して評価し，他手法との比較を通じて精度や実行速度における提案手法の有効性を確認した．

Semi-Supervised Clustering via Graph-based Projection

Tetsuya Yoshida [†1] and Kazuhiro Okatani [†1]

This paper proposes a graph-based projection approach for semi-supervised clustering based on the pairwise relations among instances. In our approach, the entire data set is represented as an edge-weighted graph with the pairwise similarities among instances. Then, in order to reflect the pairwise constraints on the clustering process, the representation is modified by contraction in graph theory and graph Laplacian in spectral graph theory. The entire data are projected onto a subspace which is constructed via the modified graph representation, and data clustering is conducted over the projected representation. The proposed approach is evaluated over several real world datasets. The results indicate that it is effective with appealing clustering performance.

†1 北海道大学大学院情報科学研究科
IST, Hokkaido University

1. はじめに

近年，ラベルありデータとラベルなしデータを活用する半教師あり学習が注目を集めている[1),2)]．この理由として，半教師あり学習では個々のデータやデータ対間の関係に対する少量のラベルありデータと大量のラベルなしデータを用いることにり，学習に必要なデータを準備する手間を抑えながら性能を格段に向上できることが挙げられる．文献[1)] では分類学習に対して半教師あり学習が PAC 学習可能であることが示され，この手法はウェブ上の求人情報の分類に対する商用サービスとしても用いられた．

他方，クラスラベルを必要としない教師なし学習としてクラスタリングの研究が行われてきた．クラスタリングとは，類似するデータは同じグループに割り当てられ，類似しないデータは異なるグループに割り当てられるように，データ全体をいくつかのグループ（クラスタ）に分割する処理である[7)]．クラスタリングを行う際には教師情報（ラベルありデータ）を必要としないが，扱うデータに対する領域知識からクラスタ割り当てに対する制約が活用できる場合もあり，その情報を用いて性能を向上させたいという要望がある．

本稿では，データ対間に対して must-link と cannot-link と呼ばれる 2 種類の制約が与えられる場合に対して，制約を反映するグラフ表現に基づく射影を用いて半教師ありクラスタリングを行う手法を提案する．データ対間の類似度が与えられた場合，類似度を重みとする辺でデータ対を繋ぐことによりデータ全体を重み付きグラフとして表現できる．データ対間での制約に対して，提案手法ではグラフ理論における縮約[6)] とスペクトルグラフ理論におけるグラフラプラシアンによる射影[3),10)] を用いることでそれぞれの制約を反映した射影表現を構築し，この表現に対してクラスタリングを行う．提案手法を高次元でスパースな表現を持つ実データに対して評価し，他手法との比較を通じて精度や実行速度における提案手法の有効性を確認した．

2. 制約に基づく半教師ありクラスタリング

以下では，$\boldsymbol{X}$でデータ集合を表記し，$|\boldsymbol{X}|$ で集合の大きさ (要素数) を表記する．

本稿では，(与えられた少量の) データ対間に対する制約の下での半教師ありクラスタリング問題を扱う．この問題は以下のように定式化される．

問題 1 (半教師ありクラスタリング). 与えられたデータ集合$\boldsymbol{X}$と制約に対して，制約を満たすデータ集合$\boldsymbol{X}$の分割（クラスターの集合）を求めよ．

様々な制約が考えられるが，本稿では must-link, cannot-link と呼ばれる 2 種類の制約を

Vol.2010-MPS-77 No.19
2010/3/5

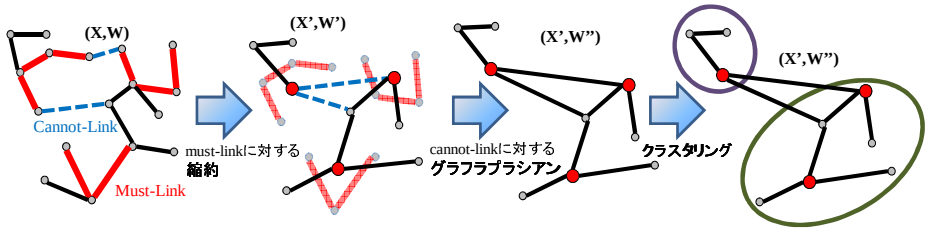

図 1　提案手法の概要
Fig. 1　Overview of graph-based projection approach.

考える[11]．

定義 1 (データ間の制約). 与えられたデータ集合$\boldsymbol{X}$と分割（クラスタ集合）$\boldsymbol{T} = \{t_1, \ldots, t_k\}$に対して，must-link 制約 $\boldsymbol{C}_{ML}$ と cannot-link 制約 $\boldsymbol{C}_{CL}$ は以下で定義される．

$$\boldsymbol{C}_{ML} = \{(x_i, x_j) | x_i, x_j \in \boldsymbol{X}, \exists t \in \boldsymbol{T}, x_i \in t \wedge x_j \in t\} \quad (1)$$

$$\boldsymbol{C}_{CL} = \{(x_i, x_j) | x_i, x_j \in \boldsymbol{X}, \exists t_a, t_b \in \boldsymbol{T}, t_a \neq t_b, x_i \in t_a \wedge x_j \in t_b\} \quad (2)$$

$\boldsymbol{C}_{ML}$ で指定されたデータ対は同じクラスタに属し，$\boldsymbol{C}_{CL}$ で指定されたデータ対は異なるクラスタに属するという制約を表現する．

3. グラフ表現に基づく半教師ありクラスタリング

3.1 準　　備

頂点集合$\boldsymbol{V}$と辺集合$\boldsymbol{E} \subset \boldsymbol{V} \times \boldsymbol{V}$ から構成されるグラフを $G(\boldsymbol{V}, \boldsymbol{E})$ と表記する．$G(\boldsymbol{V}, \boldsymbol{E})$ における辺集合$\boldsymbol{E}$は頂点集合$\boldsymbol{V}$に含まれる頂点の間の 2 項関係を表現する．

辺重み付きグラフ $G(\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{W})$ は各辺に重みが付いたグラフであり，$\boldsymbol{W}$ は重みの集合である．$|\boldsymbol{V}| = n$ の場合，重みの集合は $n \times n$ 行列 $\mathbf{W}$ で表現することができ[⋆1]，行列 $\mathbf{W}$ の第 ij 要素は頂点対 (v_i, v_j) の間の辺に対する重みを表す．本稿では重みは非負とし，辺がない頂点対間での重みは 0 とする．また，以下では無向で自己ループのない単純グラフを扱う．このため，グラフにおける重みを表現する行列 $\mathbf{W}$ は非負の要素を持つ対称行列であり，対角要素はすべて 0 である．

3.2 グラフ表現に基づくアプローチ

データ集合$\boldsymbol{X}$におけるデータ対間の類似度が与えられる場合に対し，本稿では制約付きクラスタリング問題に対するグラフ構造に基づくアプローチを提案する．類似度に基づいてデータ集合全体をデータ対間の類似度を重みとする辺重み付きグラフ $G(\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{W})$ として表現する．各データと頂点は 1 対 1 に対応するため，以下では$\boldsymbol{X}$でデータグラフにおける頂点集合も表記することとする．

定義 1 における 2 種類の制約を扱うため，提案手法では must-link 制約 $\boldsymbol{C}_{ML}$ に対してグラフ理論における縮約[6] に基づいた表現を構築する．他方，cannot-link 制約 $\boldsymbol{C}_{CL}$ を反映した最適化問題を定義し，スペクトルグラフ理論におけるグラフスペクトル[3),10)] に基づいて射影した表現を構築し，構築した表現に対してクラスタリングを行う．それぞれの詳細は 3.3 節と 3.4 節で述べる．提案手法の概要を図 1 に示す．

3.3 must-link 制約に対する縮約

式 (1) における must-link 制約 $\boldsymbol{C}_{ML}$ はデータ対が同じクラスタに属するという制約を表現する．この制約に対しては，$\boldsymbol{C}_{ML}$ に含まれる 2 対のペア (x_i, x_j)，(x_j, x_l) がある場合には，共通の x_j を介して x_i と x_l も同じクラスタに含まれるという推移律が成立する．

$\boldsymbol{C}_{ML}$ で表現される制約における推移律を扱うため，データ集合$\boldsymbol{X}$を表現するグラフ G に対して $\boldsymbol{C}_{ML}$ に基づいてグラフの縮約[6] を行う．

定義 2 (縮約). グラフ $G = (\boldsymbol{X}, \boldsymbol{E})$ の辺 e=(x_i, x_j) に対して，辺 e を新しい頂点 x_e に縮約して生成されるグラフ $G/e = (\boldsymbol{X}', \boldsymbol{E}')$ を以下で定義する．

$$\boldsymbol{X}' = (\boldsymbol{X} \backslash \{x_i, x_j\}) \cup \{x_e\} \quad (3)$$

$$\boldsymbol{E}' = \{(u, v) \in \boldsymbol{E} | \{u, v\} \cap \{x_i, x_j\} = \phi\} \cup \{(x_e, u) | (x_i, u) \in \boldsymbol{E} \backslash \{e\} \text{ or } (x_j, u) \in \boldsymbol{E} \backslash \{e\}\} \quad (4)$$

辺 e を新しい頂点 x_e に縮約することにより，頂点 x_e は辺 e=(x_i, x_j) における頂点 x_i, x_j が隣接していた全ての頂点に隣接することになる．この操作を $\boldsymbol{C}_{ML}$ に含まれる全てのデータ対に繰り返し適用することにより，must-link 制約における推移律を反映したグラフ構造を構築できる．

与えらたデータ集合$\boldsymbol{X}$に対する重みは$\boldsymbol{X}$におけるデータ対間の類似度を表現しており，この重みに基づいてデータ集合$\boldsymbol{X}$は辺重み付きグラフとして表現されていた．このため，集合 $\boldsymbol{C}_{ML}$ に含まれる各頂点対に対応する辺 e=(x_i, x_j) を縮約して新しい頂点 x_e を生成した場合には，縮約により生成されたグラフ G/e における辺の重みを定義する必要がある．ここで，must-link 制約は指定されたデータ対が両方とも同じクラスタに割り当てられるという意味でデータ対間の類似度を高めるものである．このため，$\boldsymbol{C}_{ML}$ に対応する辺を縮約したグラフ G/e では，元のグラフ G における類似度 (重み) が少なくとも保存される必要が

⋆1 太字のイタリック文字は集合，太字は行列の表記に対応する．

Vol.2010-MPS-77 No.19
2010/3/5

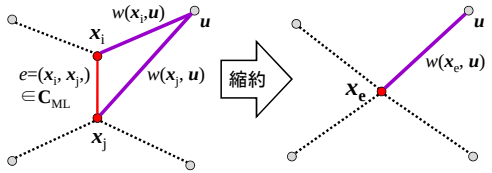

図 2 辺の縮約
Fig. 2 Contraction of an edge.

ある．

上記に基づき，$\boldsymbol{C}_{ML}$ に含まれる辺 e に対する縮約後のグラフ G/e における辺の重みを以下で定義する．

$$w(x_e,u)=\max(w(x_i,u),w(x_j,u)) \quad (x_i,u)\in \boldsymbol{E}\ or (x_j,u)\in \boldsymbol{E} \tag{5}$$

$$w(u,v)=w(u,v) \quad otherwise \tag{6}$$

縮約により生成された頂点 x_e との類似度に対しては式 (5) において関数 max を用い，また，縮約に無関係な頂点同士の重みを式 (6) により保存することにより，縮約を行った場合での辺の重みの単調非減少性を実現している．縮約操作を図 2 に示す．

集合 $\boldsymbol{C}_{ML}$ における各データ対に対して定義 2 の縮約を適用し，式 (5)，(6) による重み更新を行う．この縮約操作により更新されたグラフを $G^{'}(\boldsymbol{X}^{'},\boldsymbol{E}^{'},\boldsymbol{W}^{'})$, $n^{'}=|\boldsymbol{X}^{'}|$ と表現する．

3.4 cannot-link 制約に対するグラフラプラシアン

本稿では，cannot-link 制約を反映したクラスタリングを行うためにスペクトルクラスタリング[10] を用いる．類似するデータは同じグループに割り当て，類似しないデータは異なるグループに割り当てる，という問題を，類似したデータに類似した値を割り当て，逆に類似しないデータには異なる値を割り当てるような関数 f: $\mathcal{X}\rightarrow\mathcal{R}^{+}$ を同定する問題と捉えた場合，クラスタリングは目的関数を最小化するような関数 f を求めるという最適化問題として定式化できる．

クラスタリング処理に cannot-link 制約を反映するため，まず以下の目的関数を考える．

$$J=\frac{1}{2}\sum_{i,j\in G^{'}} w_{ij}^{'}||f_i-f_j||^2-\frac{\lambda}{2}\sum_{u,v\in C_{CL}^{'}} w_{uv}^{'}||f_u-f_v||^2 \tag{7}$$

i,j は縮約後のグラフ $G^{'}$ 上の頂点に対応し，$C_{CL}^{'}$ は $G^{'}$ における cannot-link 制約に対応する．f_i はデータ x_i に対する関数 f の値であり，$\lambda\in[0,1]$ はパラメータである．

式 (7) では，第 1 項はスペクトルグラフ理論における関数 f の滑らかさに対応し，第 2 項は $\boldsymbol{C}_{CL}$ の影響を反映した正則化項に対応する．このため，式 (7) を最小化することにより，第 2 項で $\boldsymbol{C}_{CL}$ の影響を反映しながら，類似したデータに類似した値を割り当てるような関数 f を求めることが実現される．

式 (7) の目的関数から，$\boldsymbol{C}_{CL}$ を反映した以下のグラフラプラシアン $\mathbf{L}^{''}$ を導出できる．

$$J=\sum_{i\in G^{'}}(d_i^{'}-\lambda d_i^{c})f_i^2-\sum_{i,j\in G^{'}}(w_{ij}^{'}-\lambda w_{ij}^{c})f_if_j \tag{8}$$

$$=\boldsymbol{f}^t\mathbf{D}^{''}\boldsymbol{f}-\boldsymbol{f}^t\mathbf{W}^{''}\boldsymbol{f} \tag{9}$$

$$=\boldsymbol{f}^t\mathbf{L}^{''}\boldsymbol{f} \tag{10}$$

$\boldsymbol{f}^t$ はベクトル$\boldsymbol{f}$の転置，$\cdot$ は行列の要素積を表現し，その他の項は以下で定義される．

$$(\mathbf{C}^{'})_{uv}=\begin{cases}1 & (x_u,x_v)\in C_{CL}^{'}\\ 0 & \text{otherwise}\end{cases} \tag{11}$$

$$\mathbf{W}^{c}=\mathbf{C}^{'}\cdot\mathbf{W}^{'} \tag{12}$$

$$\mathbf{W}^{''}=\mathbf{W}^{'}-\lambda\mathbf{W}^{c} \tag{13}$$

$$d_i^{'}=\sum_{j=1}^{n^{'}}w_{ij}^{'},\quad d_i^{c}=\sum_{j=1}^{n^{'}}w_{ij}^{c} \tag{14}$$

$$\mathbf{D}^{''}=diag(d_1^{''},\ldots,d_{n^{'}}^{''}), \tag{15}$$

$$d_i^{''}=d_i^{'}-\lambda d_i^{c} \tag{16}$$

$$\mathbf{L}^{''}=\mathbf{D}^{''}-\mathbf{W}^{''} \tag{17}$$

上記は縮約後のグラフ $G^{'}$ を式 (13) で定義される重み $\mathbf{W}^{''}$ を持つグラフ $G^{''}$ に変換することに対応する．このため，データ集合に対して提案手法はデータ対間の類似度と制約 ($\boldsymbol{C}_{ML}$ と $\boldsymbol{C}_{CL}$) を反映したグラフ表現を構築することになる（図 1 参照）．

スペクトルクラスタリングにおいては生成するクラスタ相互のバランスを考慮することが重要になることが知られており[10]，通常は正規化した目的関数が用いられる．本稿でも，式 (14)，(17) に基づき，式 (7) での目的関数を正規化した以下の目的関数を考える．

$$J_{sym}=\frac{1}{2}\sum_{i,j}w_{ij}^{''}||\frac{f_i}{\sqrt{d_i^{''}}}-\frac{f_j}{\sqrt{d_j^{''}}}||^2 \tag{18}$$

図 3　アルゴリズム GBSSC
Fig. 3　Algorithm GBSSC

GBSSC(G, $\boldsymbol{C}_{ML}$, $\boldsymbol{C}_{CL}$, l, k)
Require: $G(\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{W})$; *//an edge-weighted graph*
Require: $\boldsymbol{C}_{ML}$; *//must-link constraints*
Require: $\boldsymbol{C}_{CL}$; *//cannot-link constraints*
Require: l; *//the dimensions of the subspace*
Require: k; *//the number of clusters*

1: **for** each $e \in C_{ML}$ **do**
2:　contract e and create contracted graph G/e;
3: **end for**
　// *Let* $G^{'}(\boldsymbol{X}^{'}, \boldsymbol{E}^{'}, \boldsymbol{W}^{'})$ *be the contracted graph.*
4: create $\mathbf{C}^{'}_{uv}, \mathbf{W}^{c}, \mathbf{W}^{''}, \mathbf{D}^{''}$ as eqs.(11) $\sim$ (15).
5: $\mathbf{L}^{''}_{sym} = \mathbf{I} - \mathbf{D}^{''-\frac{1}{2}}\mathbf{W}^{''}\mathbf{D}^{''-\frac{1}{2}}$
6: Find l eigenvectors $\mathbf{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$ for $\mathbf{L}^{''}_{sym}$, with the smallest non-zero eigenvalues.
7: Conduct clustering of data which are represented as $\mathbf{H}$ and construct clusters.
8: **return** clusters

式 (18) における目的関数 J_{sym} を最小化することは一般化固有値問題 $\mathbf{L}^{''}\boldsymbol{h} = \alpha\mathbf{D}^{''}\boldsymbol{h}$ に対する固有ベクトルを求めることに対応する．ここで，$\boldsymbol{h}$は固有ベクトル，α は固有値に対応する．正の固有値に対応する一般化固有ベクトルの集合 $\mathbf{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$ を固有値の昇順に求めることにより，これらのベクトルで張られる部分空間に与えられたデータ集合を射影した表現が構築される．

3.5　アルゴリズム

提案するアルゴリズム GBSSC(graph-based semi-supervised clustering) を図 3 に示す．行 1 から行 3 では $\boldsymbol{C}_{ML}$ に対応する辺を縮約して縮約後のグラフ $G^{'}$ を構築する．行 4 から行 6 では式 (18) における目的関数 J_{sym} の最小化を行う．アルゴリズムにおいては J_{sym} は 5 行目の正規化グラフラプラシアン $\mathbf{L}^{''}_{sym}$ として表現される．6 行目で $\mathbf{L}^{''}_{sym}$ に対する固有ベクトルの集合 $\mathbf{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$ を求め，生成した表現 $\mathbf{H}$ に対して 7 行目でクラスタリングを行い，クラスタを生成する．現状ではクラスタリングには skmeans [⋆1][14] を用いている．

⋆1 skmeans は高次元スパース表現に対する標準的な手法である

表 1　20 Newsgroup に対するデータセット
Table 1　Datasets from 20 Newsgroup dataset

データセット	含まれるグループ名
Multi5	comp.graphics, rec.motorcycles,rec.sport.baseball, sci.space,talk.politics.mideast
Multi10	alt.atheism, comp.sys.mac.hardware,misc.forsale, rec.autos,rec.sport.hockey, sci.crypt,sci.med,sci.electronics,sci.space,talk.politics.guns
Multi15	alt.atheism, comp.graphics, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.guns, talk.politics.mideast, talk.politics.misc

4. 評　　価

4.1 実験設定

4.1.1　対象データ

先行研究[5),9)] に基づき，提案手法を 20 ニュースグループ (以下，20NG) [⋆2]に対して評価した．これらは単語の頻度に基づくベクトル空間モデルで表現された文書データであるため，文書クラスタリングを行うことに対応する．文書クラスタリングとは文書集合$\boldsymbol{X}$=$\{x_1, \ldots, x_n\}$をクラスタ集合$\boldsymbol{T}$に分割する問題である．一般に文書に含まれる単語数は膨大であり，また文書ごとに含まれる単語が異なることが多いため，高次元スパース表現なデータをクラスタリングすることに対応する．

20NG に対して 5 クラスタ，10 クラスタ，15 クラスタからなる 3 つの母集団を設定し，各母集団に含まれるクラスタからそれぞれ 50 個ずつの文書を非復元抽出してデータセットを作成した．各母集団に含まれるニュースグループを表 1 に示す．各母集団に対して 10 個ずつ，計 30 個のデータセットを作成した．各データセットごとに porter stemmer [⋆3] を用いて stemming を行い，MontyTagger [⋆4] を用いて名詞を抽出し，stop word を除去して相互情報量で上位 2,000 語の単語を選択した．

4.1.2　評価尺度

上記のデータは，各データ（ここでは文書）ごとに真のクラスタが既知である．各データセットに対して，各データに対する真のクラスタと割り当てられたクラスタに基づいて正規化相互情報量 (NMI) を評価した．

⋆2 http://people.csail.mit.edu/ jrennie/20Newsgroups/．本稿では 20news-18828 を使用した．
⋆3 http://www.tartarus.org/ martin/PorterStemmer
⋆4 http://web.media.mit.edu/ hugo/montytagger

真のクラスタと割り当てられたクラスタに対応する確率変数を T, $\hat{T}$ とすると，正規化相互情報量 (NMI) は以下で定義される．

$$NMI = \frac{I(\hat{T};T)}{(H(\hat{T})+H(T))/2} \quad (\in [0,1]) \tag{19}$$

$H(T)$ はシャノン情報量である．NMI における正規化には様々な手法があるが，本稿では平均による正規化とした．NMI が大きいほど真のクラスタでのデータ割り当てに合致することを示すため，クラスタ割り当ての正当性（精度）に対応する．

また，実験で比較した手法はすべてまずクラスタリングに用いる表現を構築し，構築した表現に対して標準的なクラスタリング手法 (kmeans など) を適用する．このため，実行速度の評価として表現の構築に要する CPU 時間 (秒) を計測した．実験は Windows Vista, Intel Core2 Quad Q8200 2.33GHz, 2G メモリの計算機で行った．

4.1.3 比 較 手 法

提案手法を，SCREEN[9], PCP[8], と比較した．比較手法は全て分割的クラスタリングを行うものであるためクラスタ数 k は与えられると仮定した．

4.1.4 実験パラメータ

定義 1 におけるデータ対間の制約に対するパラメータは 1) 制約数，2) 制約を指定するデータ対，である．2) に関しては，データ対の非復元抽出により制約を生成した．このため，主要なパラメータは $\boldsymbol{C}_{ML}$ と $\boldsymbol{C}_{CL}$ に対する制約数である．以下では $|\boldsymbol{C}_{ML}| = |\boldsymbol{C}_{CL}|$ とし，制約数 $|\boldsymbol{C}_{ML}|$ を変えて実験を行った．

データ対間の距離尺度としては，文献[9] に従って各データセットに対する p 次元表現 (p は属性数) におけるユークリッド距離を用いた．各データ $\boldsymbol{x} \in \mathcal{R}^p$ の長さを $\boldsymbol{x}^t\boldsymbol{x} = 1$ と正規化し，文書処理で標準的に用いられるコサイン類似度により類似度を定義した．

提案手法と PCP では，上記の類似度に基づいて重み付きグラフを構築する．提案手法では完全グラフを構築したが，完全グラフに対して PCP を適用すると非常に精度が悪かった．このため，文献[8] に従って PCP に対しては各データごとに類似度が上位 m 個の近傍データから m-近傍グラフを構築し，文献[8] に従って近傍数 m=10 とした．

提案手法と SCREEN では，写像先の部分空間の次元数 l を指定する必要がある．次元数 l も結果に影響を及ぼすが，次元数 l=クラスタ数 k とした．提案手法における式 (7) でのパラメータ λ は 0.5 とした．

SCREEN では線形写像を求める際に $\boldsymbol{C}_{CL}$ で指定されたデータに対して $p \times p$ の共分散行列を構築する．高次元データ (次元数 p が大きい場合) では行列のサイズが大きくなってしまうため，実際には高次元データに適用しにくいという課題がある．この問題に対処するため，文献[9] ではまず主成分分析 (PCA) を用いて次元圧縮して低次元表現を求め，この低次元表現に対して SCREEN を適用している．文献[9] に従い，寄与率の観点から上位 100 個の主成分を選択して低次元表現を生成した．

4.1.5 実 験 手 順

それぞれの制約数に対して制約 $\boldsymbol{C}_{ML}$ と $\boldsymbol{C}_{CL}$ を生成し，生成した制約のもとでクラスタリングを行った．クラスタ割り当ては初期化の影響を受けるため，生成した制約のもとでクラスタリングを 10 回行った．更に，この処理を制約数ごとに 10 回繰り返した．このため，各データセットごとに制約数に対して 100 回試行を行い，その平均値を求めた．

4.2 実データに対する評価

実験結果において横軸は制約数に対応する．縦軸は式 (19) で定義した NMI，あるいは 1 回の試行あたりの CPU 時間 (秒) の平均値に対応する．図中の凡例において，+PCA は主成分分析を用いて低次元表現を生成し，その表現に対して手法を適用した結果を示す．GBSSC+PCP は 3.3 節で提案した手法で must-link 制約に対して縮約を行い，cannot-link 制約に対しては m-近傍グラフを構築して PCP を適用した場合を示す．

表 1 に示した 20NG に対する結果 (NMI, 実行速度) を図 4，5 に示す[⋆1]．それぞれの図は各母集団ごとに対する 10 データセットの平均値である．

文書クラスタリングは高次元スパースデータをクラスタリングすることに対応するが，図 4 より，部分空間の次元数 l=クラスタ数 k とした場合 [⋆2] に提案手法 (GBSSC, 赤色) は他手法を上回る性能を示した．Multi5 に対しては制約数が増加するにつれ PCP(緑色) は GBSSC に近づき，制約数=100 ではほぼ同程度の性能を示したが，図 5 に示すように提案手法 (GBSSC) は 2 桁 (100 倍) 以上高速であった．

提案手法を用いて must-link 制約に対して縮約し，cannot-link 制約に対して PCP を適用した場合 (GBSSC+PCP，ピンク) は，精度 (NMI) は PCP とほぼ同程度あった．しかし，must-link 制約の縮約により PCP と比べて 1 桁 (10 倍) 程度高速であったが，提案手法 (GBSSC) のほうが更に 1 桁 (10 倍) 程度高速であった．

⋆1 図 5 に示すように PCP は非常に実行時間がかかる．このため Multi15 に対してのみ PCP では 3 データセットの平均となっているが，発表時には 10 データセットに対する平均を報告する予定である．
⋆2 本稿では次元数 l に対するパラメータチューニングは行っていない．

Vol.2010-MPS-77 No.19
2010/3/5

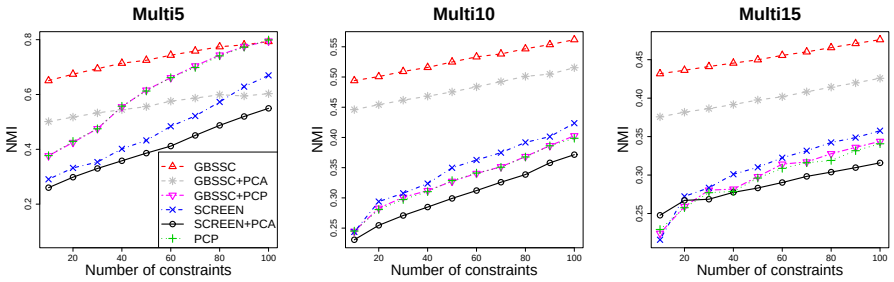

図 4 20Newsgropu に対する結果 (NMI)
Fig. 4 Results on 20 Newsgroup datasets (NMI)

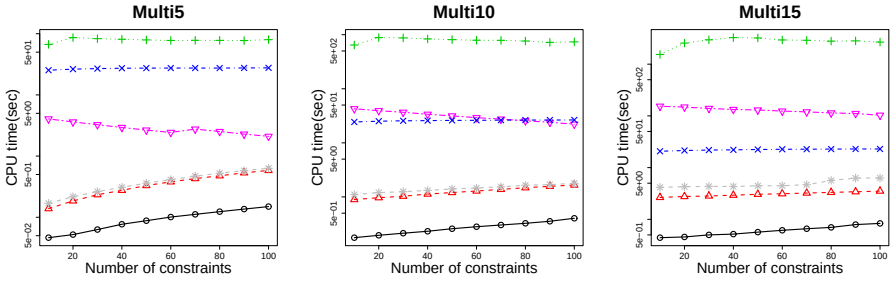

図 5 20Newsgropu に対する結果 (CPU 時間)
Fig. 5 Results on 20 Newsgroup datasets (CPU time)

主成分分析の使用は SCREEN(青色) の高速化には役だったが[⋆1]，提案手法に対しては効果がなかった（GBSSC+PCA, 灰色）．また，精度 (NMI) はともに悪化した．このため，提案手法では主成分分析による低次元表現の生成は不要であると言える．

4.3 考　　察

4.2 節での結果より，文書データなどの高次元スパースなデータに対して精度 (NMI) および実行時間の観点から提案手法の有効性を確認した．SCREEN[9] に対しては，提案手法は精度および計算速度の観点で大きく上回る性能を示した．PCP[8] に対しては，精度に対しては劣る場合もあったが 2 桁 (100 倍) 以上高速であった．上記より，提案手法は半教師ありクラスタリングを行うための効果的な手法であると考えられる．

⋆1 SCREEN+PCA(黒色) は SCREEN(青色) より 3 桁 (1000 倍) 程度高速であった．

5. おわりに

本稿では，must-link と cannot-link と呼ばれる制約が与えられる場合に対して制約を反映するグラフ表現に基づく射影を用いて半教師ありクラスタリングを行う手法を提案した．提案手法ではデータ対間の類似度に基づいてデータ全体を辺重み付きグラフとして表現し，グラフ理論における縮約とグラフラプラシアンによる射影用いてそれぞれの制約を反映した射影表現を構築し，構築した射影表現に対してクラスタリングを行う．提案手法を高次元スパースな表現を持つ実データに対して評価し，他手法との比較を通じて精度や実行速度における提案手法の有効性を確認した．今後は画像データなどの他の実データに対しても評価を行い，提案手法を改良していく予定である．

謝辞 本研究の一部は文部科学省科研費 (No. 20500123) の補助による．

参　考　文　献

1) Blum, A. and Mitchell, T.: Combining Labeled and Unlabeled Data with Co-Training, *Proc. of COLT-98*, pp.92–100 (1998).
2) Chapelle, O., Schölkopf, B. and Zien, A.(eds.): *Semi-Supervised Learning*, MIT Press (2006).
3) Chung, F.: *Spectral Graph Theory*, American Mathematical Society (1997).
4) Dhillon, J. and Modha, D.: Concept decompositions for lage sparse text data using clustering, *Machine Learning*, Vol.42, pp.143–175 (2001).
5) Dhillon, J., Mallela, S. and Modha, D.: Information-theoretic co-clustering, *Proc. of KDD'03*, pp.89–98 (2003).
6) Diestel, R.: *Graph Theory*, Springer (2006).
7) Jain, A., Murty, M. and P.J., F.: Data Clustering: A Review, *ACM Computing Surveys*, Vol.31, pp.264–323 (1999).
8) Li, Z., Liu, J. and Tang, X.: Pairwise constraint propagation by semidefinite programming for semi-supervised classification, *Proc. of ICML-08*, pp.576–583 (2008).
9) Tang, W., Xiong, H., Zhong, S. and Wu, J.: Enhancing Semi-Supervised Clustering : A Feature Projection Perspective, *Proc. of KDD'07*, pp.707–716 (2007).
10) von Luxburg, U.: A Tutorial on Spectral Clustering, *Statistics and Computing*, Vol.17, No.4, pp.395–416 (2007).
11) Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S.: Constrained K-means Clustering with Background Knowledge, *Proc. of ICML'01*, pp.577–584 (2001).