情報処理学会論文誌　数理モデル化と応用　Vol. 4　No. 1　62–71 (Jan. 2011)

制約を反映するグラフ構造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，制約を反映させ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させ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
Graduate School of Information Science and Technology, Hokkaido University

1. はじめに

計算機の処理装置や記憶容量の高性能化・低価格化などにともない，web ページなどの膨大なデータを蓄積して処理することが可能になった．このため，大量のデータから自動的に学習させるという機械学習の手法が鋭意研究開発されている．たとえば分類器学習を用いたスパムメールの自動フィルタリングなども実用に供されている．しかし，分類器の構築にはクラスラベルが付随するデータ（ラベルありデータ）を準備する必要がある．一般にはラベルありデータの量が多いほど高精度な分類が可能になるが，個々のデータの内容に基づいてクラスラベルを付与することの負荷が高いという課題がある．

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[4),5)]．この理由として，半教師あり学習では個々のデータ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させることができる点があげられる．文献 4) 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．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当てへの制約（たとえば，あるデータ対が同じクラスタに属する，あるいは属さないという制約）が活用できる場合もあり，これを活用して性能を向上させたいという要望がある．

クラスタ割当てへの制約を教師情報ととらえることで，与えられた制約の下で半教師ありクラスタリングを行う様々な研究が行われている[1),15),16),20)]．文献 18) は kmeans 法[13)] を用いる際に制約充足を各データごとに確認し，制約が充足されるクラスタに限定して割り当てる手法を提案した．文献 16) は高次元データに対して線形写像に基づく次元縮約を用いる手法を提案し，他手法との比較を通じて有効性を示した．しかし，この手法ではクラスタ割当てへの制約は活用されたが，データ対間での関係は明示的には用いられていなかった．

本稿では，クラスタ割当てに対して 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，制約を反映させ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．データ対の関係が類似度として与えられた場合，類似

度を重みとする辺でデータどうしをつなぐことにより，データ全体を重み付きグラフとして表現できる．さらに，クラスタ割当てへの制約を反映するため，グラフ理論における縮約[9]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[2),6),17)]を用いてそれぞれの制約を反映させた射影表現を構築し，射影表現に対してクラスタリングを行う．

データ対間の類似度に基づいてデータ全体を重み付きグラフとして表現し，クラスタ割当てへの制約を反映させた射影表現を構築することにより，データ対間の関係（類似度と制約）に基づいて半教師ありクラスタリングを効果的に行うことが可能となる．提案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．

2 章で制約に基づく半教師ありクラスタリングの概略を説明し，3 章で提案手法の詳細について説明する．4 章で提案手法の評価を述べ，5 章でまとめと今後の展望を述べる．

2. 制約に基づく半教師ありクラスタリング

クラスタリングには大きく分けて階層的クラスタリングと分割的クラスタリングのアプローチがある．前者ではデンドログラムと呼ばれる木構造を構築してデータ集合を分割し，木構造における部分木によりそれぞれクラスタを表現する．他方，後者ではクラスタ数などを設定し，各データをクラスタに割り当てることでデータ集合を分割する．

2.1 制約に基づくクラスタリング

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

本稿では，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と制約の下でデータ集合の分割（クラスタの集合）を求めよ．

様々な制約が考えられるが，本稿では文献 18) に従い must-link，cannot-link と呼ばれる 2 種類の制約を考える．

定義 1（データ対への制約）．与えられたデータ集合Xと分割（クラスタ集合）$T = \{t_1, \ldots, t_k\}$ に対して，must-link C_{ML} と cannot-link C_{CL} はそれぞれデータ対の集合であり，以下を満たすものである．

$$(x_i, x_j) \in C_{ML} \Rightarrow \exists t \in T\ (x_i \in t \wedge x_j \in t) \quad (1)$$

$$(x_i, x_j) \in C_{CL} \Rightarrow t_a, t_b \in T, t_a \neq t_b\ (x_i \in t_a \wedge x_j \in t_b) \quad (2)$$

C_{ML} で指定されたデータ対は同じクラスタに属し，C_{CL} で指定されたデータ対はそれぞれ異なるクラスタに属するという制約を表現する．なお，クラスタ割当てへの制約はすべてのデータ対間に定義されるとは限らず，半教師ありクラスタリングではできるだけ少量の制約を用いて性能を向上させることが望ましい．

2.2 関 連 研 究

分割的クラスタリングに対しては様々な手法が提案されているが，グラフ理論に基づいてデータ集合をグラフ構造の観点からとらえるアプローチがある．このアプローチでは，類似度（あるいは非類似度）に基づきグラフの中から組合せ論的な構造を探索する．これまで，グラフ彩色に基づく手法[10),12)] やグラフカットに基づく手法[17)] などが提案されている．本稿のアプローチはグラフ構造に基づく分割的クラスタリングに位置づけられる．

半教師ありクラスタリングには制約に基づく手法，距離に基づく手法，両者を統合する手法，という 3 つのアプローチがある．制約に基づく手法では指定されたデータ間の制約をクラスタリング処理に反映させるものが多い[16),18)]．距離に基づく手法では，距離学習を用いて指定されたデータ間の制約からクラスタリング処理で用いる距離尺度を学習して利用するアプローチがある[15),20)]．また，上記の 2 つを確率的な枠組みに基づいて統合する手法も提案されている[1)]．本稿での提案手法は制約に基づく手法に位置づけられる．

定義 1 で述べた制約の下での半教師ありクラスタリング問題に対し，文献 18) は kmeans 法[13)] に基づく COP-kmeans を提案した．COP-kmeans ではクラスタ中心との最短距離規範に基づいて各データのクラスタ割当てを決定する際，与えられた制約の充足を各データごとに確認し，制約が充足されるクラスタにデータを割り当てる．

SCREEN [16)] は must-link に基づいて与えられたデータ表現を変換し，cannot-link が指定されたデータ間での分散が最大化される部分空間への線形写像を求め，線形写像による射影表現に対してクラスタリングを行う．射影表現に対してクラスタリングする際には，COP-kmeans における制約充足を cannot-link が指定されたデータ対に限定し，C_{CL} に含まれるデータ対が極力離れるように改良した PCkmeans を用いる．

PCP [15)] は，与えられた制約の下で半正定値計画問題を解いて写像先の空間におけるデータ間の類似度（カーネル行列）を求めてクラスタリングを行う手法である．写像自体や，写像先の空間における各データの明示的な表現は求められないが，与えられたデータ集合に対するカーネル行列が定まるため，カーネル kmeans を用いてクラスタリングを行う．

スペクトルクラスタリング[17)] とは，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ベクトルを行列の固有値・固有ベクトルに基づいて近似的に求める手法である．文献 3) はクラスタ割当てへの制約を $\{0, 1, -1\}$ を要素

とする行列として表現し，この行列に基づいて制約付きスペクトルクラスタリングを行列の固有値問題として定式化したが，cannot-link は 2 クラスだけに限定されていた．文献 14) は負の重みを持つグラフに対するグラフラプラシアンを提案し，2 次元平面への埋込みの例を示しているが，制約付きクラスタリングとしての評価は行っていない．

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

辺重み付きグラフ $G(V,E,W)$ は各辺に重みが付いたグラフであり，Wは辺への重み割当て関数である．$|V| = n$ の場合，重み割当ては $n \times n$ 行列 $\mathbf{W}$ で表現でき，$\mathbf{W}$ 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 $[0,1]$ の実数とし，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，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，本稿ではデータ全体を類似度を重みとする辺重み付きグラフ $G(V,E,W)$ として表現して，半教師ありクラスタリング問題に対してグラフ構造に基づくアプローチを提案する．各データと頂点は 1 対 1 に対応するため，以下ではXでグラフにおける頂点集合も表記する．対 (x_i, x_j) に対する重み w_{ij} はデータ間の類似度に対応し，重みが大きいほど x_i, x_j が類似していることになる．

本稿では，データ対の関係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[2),17)] を用いる．類似するデータを同じグループに，類似しないデータを異なるグループに割り当てる，というクラスタリングの処理を，類似したデータには類似した値を，類似しないデータには異なる値を割り当てるベクトル $\boldsymbol{h}$ を同定する問題ととらえた場合，クラスタリングを以下を満たすベクトル $\boldsymbol{h}$ を求める問題として定式化できる[17)]．

$$\mathbf{D} = diag(d_1, \ldots, d_n) \quad \left(\text{ただし} \quad d_i = \sum_{j=1}^{n} w_{ij} \right) \tag{3}$$

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{4}$$

$$\boldsymbol{h} = \arg\min_{\boldsymbol{h}} \boldsymbol{h}^t \mathbf{L} \boldsymbol{h} \tag{5}$$

式 (3) での $\mathbf{D}$ は $d_1, \ldots, d_n$ を要素とする対角行列であり，$\boldsymbol{h}^t$ は $\boldsymbol{h}$ の転置ベクトルを表す．

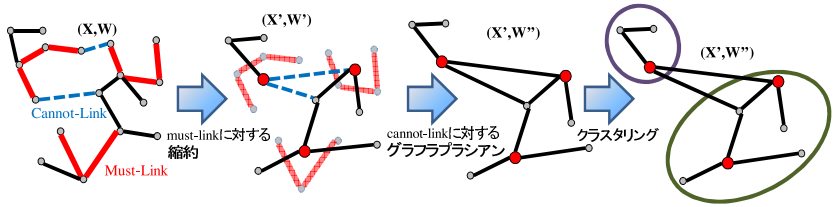

図 1　提案手法の概要
Fig. 1　Overview of graph-based projection approach.

式 (4) の行列 $\mathbf{L}$ はグラフラプラシアンと呼ばれる[6),17)]．また，スペクトルクラスタリングにおいては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，通常は $\mathbf{L}$ を正規化したものが用いられる．本稿でも，式 (5) でベクトル $\boldsymbol{h}$ を求める際，式 (3) での行列 $\mathbf{D}$ に基づいて正規化した $\mathbf{L}_{sym} = \mathbf{D}^{-\frac{1}{2}}\ \mathbf{L}\ \mathbf{D}^{-\frac{1}{2}}$ に対するベクトルを求めることとする．このベクトルは一般化固有値問題 $\mathbf{L}\ \boldsymbol{h} = \alpha\ \mathbf{D}\ \boldsymbol{h}$ に対する一般化固有ベクトルであり，α は固有値に対応する．

さらに，上記を複数のベクトルに拡張した場合，以下の目的関数を最小化する行列 $\mathbf{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$ を求める問題として定式化できる．

$$J_1 = tr(\mathbf{H}^t \mathbf{L} \mathbf{H}) \tag{6}$$
$$s.t. \quad \mathbf{H}^t \mathbf{D} \mathbf{H} = \mathbf{I}$$

式 (7) での tr は行列のトレースを表し，$\mathbf{I}$ は単位行列を表す．正の固有値に対応する一般化固有ベクトルの集合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データ集合を射影した表現が構築される．$\mathbf{H}$ は $n \times l$ の実行列であり（$n = |X|$ はデータ数），データ全体を l 次元の部分空間に射影した表現[⋆1]に対応する[2)]．最後に，構築した表現に対してクラスタリング手法を適用してクラスタを生成する[6),17)]．適用するクラスタリング手法として kmeans 法などがよく使用される．

定義 1 における 2 種類の制約を扱うため，提案法では must-link に対してグラフ理論における縮約[9)] に基づいたグラフ構造を構築する．他方，cannot-link を反映させるために，式 (6) に対して cannot-link に基づく正則化項を追加した最適化問題を定義する．提案法の概要を**図 1** に示す．本稿では must-link は縮約を用いるために hard 制約として扱うが，cannot-link は正則化に基づく射影表現を用いるため soft 制約として扱うことになる．

⋆1 $\mathbf{H}$ における行ベクトルが各データの表現に対応する．

3.3　must-link に対する縮約

式 (1) における must-link は対 (x_i,x_j) で指定される x_i，x_j 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，C_{ML} に含まれる 2 つの対 (x_i,x_j)，(x_j,x_k) がある場合，共通の x_j を介して x_i と x_k も同じクラスタに含まれるという推移律が成立する．

must-link における推移律を扱うため，データ集合Xを表現するグラフ G に対して C_{ML} に基づいてグラフの縮約[9] を行う．

定義 2（縮約）．グラフ $G(X,\,E)$ 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 $G/e=G'(X',E')$ を以下で定義する．

$$X'=(X\backslash\{x_i,x_j\})\cup\{x_e\} \tag{7}$$

$$E'=\{(u,v)\in E|\{u,v\}\cap\{x_i,x_j\}=\phi\} \cup\{(x_e,u)|(x_i,u)\in E\backslash\{e\}\ or\ (x_j,u)\in E\backslash\{e\}\} \tag{8}$$

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 が隣接していたすべての頂点に隣接することになる．この操作を C_{ML} に含まれるすべてのデータ対に繰り返し適用することにより，must-link における推移律を反映したグラフ構造を構築する．縮約操作を**図 2** に示す．

データ集合に対する重み $\mathbf{W}$ は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全体が辺重み付きグラフ G として表現されていた．集合 C_{ML} に含まれる辺 $e=(x_i,x_j)$ を縮約して新しい頂点 x_e を生成した場合，縮約後のグラフ G/e における辺の重みを定義する必要がある．その際，たとえば図 2 で縮約前の重み $w(x_i,u)$ の下では x_i，u が同じクラスタに割り当てられる場合であっても，縮約後の重み $w(v,u)$ がそれより小さくなると異なるクラスタに割り当てられやすくなってしまう．

本稿では C_{ML} に含まれるデータ対に対応する辺 e を縮約したグラフ G/e における辺の重みを以下で定義する．

$$w'(v,u)=\begin{cases}\max\{w(x_i,u),w(x_j,u)\} & \text{if}\ \ v=x_e\ \text{(contracted vertex)},\\ & (x_i,u)\in E,\ (x_j,u)\in E\\ w(v,u) & \text{otherwise}\end{cases} \tag{9}$$

縮約により生成された頂点 x_e との類似度に対して関数 max を用い，縮約に無関係なデータ対には元の重みを用いることで辺の重みの単調非減少性が保障される．集合 C_{ML} における各データ対に縮約を適用し，式 (9) により重みを定義する．この操作により構築されたグラフを $G^{'}(X^{'},E^{'},W^{'})$, $n^{'}=|X^{'}|$ と表現する．また，G' における重み（類似度）を行列として表現したものを $\mathbf{W}$' と表記する．

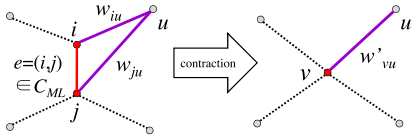

図 2　辺の縮約
Fig. 2　Contraction of an edge.

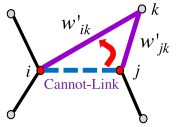

図 3　制約伝播
Fig. 3　Constraint propagation.

3.4　cannot-link に対する正則化グラフラプラシアン

3.3 節で述べた縮約操作により構築したグラフ G' に基づき，cannot-link を反映した射影表現を構築するために，式 (7) を cannot-link に基づいて正則化した以下の目的関数を最小化する行列 $\mathbf{H}$ を求める．

$$J_2=tr(\mathbf{H}^t\mathbf{L}\mathbf{H})+\lambda\ tr(\mathbf{H}^t\mathbf{S}\mathbf{H}) \tag{10}$$

$$s.t.\ \ \mathbf{H}^t\mathbf{D}\mathbf{H}=\mathbf{I}$$

ここで，行列 $\mathbf{S}$ は cannot-link に基づいて定義される行列であり，実数 λ はパラメータである．3.2 節で述べたように $\mathbf{H}^t$ は射影表現に対応し，$\mathbf{H}\mathbf{H}^t$ はそのグラム行列である．行列 $\mathbf{S}$ を用いて cannot-link に関連するデータ対の内積を $\mathbf{S}$ で重み付けして取り出すことにより，式 (10) の第 2 項が cannot-link を反映した正則化項となる．

制約として指定されたデータ対に関する情報を正則化項として用いることは他の手法でも用いられていたが，従来のアプローチでは行列 $\mathbf{S}$ を

$$s_{ij}=\begin{cases}-1 & \text{if}\ \ (x_i,x_j)\in C_{ML}\\ 1 & \text{if}\ \ (x_i,x_j)\in C_{CL}\\ 0 & \text{otherwise}\end{cases} \tag{11}$$

と設定していた[3),11),19)]．式 (11) では制約として指定されたデータ対のみが考慮され，他のデータ対は正則化に用いられないため，制約の効果が限定的であるという課題がある．

上記の課題に対し，**図 3** に示すようにデータ対 (x_i,x_j) に関する制約を頂点 x_i，x_j に接続する他の辺（データ対）にも伝播させて正則化に用いることを提案する．なお，本稿では 3.3 節の手法で must-link を反映させてグラフ G' を構築するため，以下では cannot-link に基づいて行列 $\mathbf{S}$ を定義する．$(x_i,x_j)\in C_{CL}$ である頂点 x_j に接続する頂点 x_k に関して $(x_i,x_k)\notin C_{CL}$ の場合，従来のように $s_{ik}=0$ とするのではなく，G' における重み

図 4　アルゴリズム GBSSC
Fig. 4　Algorithm GBSSC.

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

1: **for** each $e \in C_{ML}$ **do**
2:　contract e and create contracted graph
3: **end for**
　// Let $G'(X', E', W')$ *be the contracted graph.*
4: create $\mathbf{S}$ as defined in eq.(12) *//utilize* C_{CL}
5: create $\mathbf{D}$, $\mathbf{L}$ as eqs.(3), (4) for G'
6: find $\mathbf{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$ for eq.(10) with $\mathbf{S}$ in eq.(12).
7: Conduct data clustering w.r.t. $\mathbf{H}$ and construct clusters.
8: **return** clusters

$w'_{ik}, w'_{jk} \in [0,1]$ を反映させて $s_{ik} = (1 - w'_{ik})w'_{jk}$ とすることを考える．たとえば $w'_{jk} = 1$ かつ $w'_{ik} = 0$ の場合，グラフ G' においては頂点 x_j，x_k は非常に類似しており，逆に頂点 x_i，x_k はまったく似ていないことに対応する．この場合は s_{ik}=1 となり，対 (x_i, x_j) に関する制約を対 (x_i, x_k) にも伝播させて正則化に用いることになる．他方，$w'_{jk} = 0$ で頂点 x_j，x_k がまったく似ていない場合には $s_{ik} = 0$ となり，対 (x_i, x_j) に関する制約は対 (x_i, x_k) に伝播せず，この対は正則化に用いられないことになる．

対 (x_i, x_k) に対する s_{ik} を定義する際，一般には頂点 x_i に関して複数のデータ対が制約として指定される場合があり，また，頂点 x_k に関しても制約が指定される場合がある．本稿では平均により複数の制約を集約することとし，以下で行列 $\mathbf{S}$ を定義する．

$$s_{ik} = \begin{cases} 1 & \text{if } (x_i, x_k) \in C_{CL} \\ \frac{1}{m_{ik}} \left\{ \sum_{(x_i,x_j)\in C_{CL}} (1 - w'_{ik})w'_{jk} + \sum_{(x_k,x_j)\in C_{CL}} (1 - w'_{ik})w'_{ji} \right\} & \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

式 (12) の 2 行目は $((x_i, x_k) \notin C_{CL}) \wedge \{((x_i, x_j) \in C_{CL}) \vee ((x_k, x_j) \in C_{CL})\}$ である場合に対応し，m_{ik} は対 (x_i, x_k) 影響をおよぼす制約数である．式 (12) はデータ対の関係（類似度）を反映して制約をその近傍のデータ対にも拡張して正則化に用いることになる．

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

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 4** に示す．行 1 から行 3 では 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 で制約を反映した行列 $\mathbf{S}$ を式 (12) に従って構築し，行 5，6 で式 (10) を最小化する $\mathbf{H}$ を求め，行 7 でクラスタリングを行う．現状ではクラスタリングには skmeans [7] を用いている．

4. 評　　価

4.1　実 験 設 定

4.1.1　対象データ

先行研究[8),16)] に基づき，提案手法を 20 ニュースグループ（以下，20NG と表記）[⋆1]，TREC データセット[⋆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$X = \{x_1, \ldots, x_n\}$ をクラスタ集合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

⋆1 http://people.csail.mit.edu/~jrennie/20Newsgroups/．本稿では 20news-18828 を使用した．
⋆2 http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download
⋆3 http://www.tartarus.org/~martin/PorterStemmer

表 2　TREC データ
Table 2　TREC datasets.

dataset	# attr.	#clusters	#data
tr11	6429	9	414
tr12	5804	8	313
tr23	5832	6	204
tr31	10128	7	927
tr41	7454	10	878
tr45	8261	10	690

いて stemming を行い，MontyTagger[⋆1]を用いて品詞に分解し，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．

TREC データセットに対しては，文献 16) に従って**表 2** に示す 6 つのデータセットを使用した．これらのデータセットはすでに前処理済みであるため個別の前処理などは行わなかった．

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．

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．

$$NMI = \frac{I(\hat{T};T)}{(H(\hat{T})+H(T))/2} \quad (\in [0,1]) \tag{13}$$

$H(\cdot)$ はシャノン情報量であり，$I(\cdot;\cdot)$ は相互情報量である．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当てに合致することを示すため，クラスタ割当ての正当性（精度）に対応する．

また，実験で比較した手法（skmeans を除く）ではまずクラスタリングに用いる各データの表現を構築し，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生成に要する CPU 時間（秒）を計測した．実験は Debian/GNU Linux，Intel Xeon W5590，36 G メモリの計算機で行った．

⋆1 http://web.media.mit.edu/~hugo/montytagger

表 3　skmeans の結果（NMI）
Table 3　Result of skmeans (NMI).

	Multi5	Multi10	Multi15	tr11	tr12	tr23	tr31	tr41	tr45
NMI	0.380	0.333	0.333	0.574	0.500	0.283	0.407	0.530	0.510

4.1.3　比 較 手 法

提案法を 2.2 節で述べた SCREEN[16]，PCP[15]，skmeans[7] と比較した．比較手法はすべ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（以下では k と表記）は与えられると仮定した．skmeans は高次元スパースデータに対する標準的な手法であり，制約を使用しない場合のベースラインとして評価した．skmeans に対する結果（NMI）を**表 3** に示す．4.3 節で示すように，制約を用いて半教師ありクラスタリングを行うことにより提案法はすべてのデータセットにおいて skmeans を上回る性能を示した．

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 C_{ML} と C_{CL} に対する制約数である．以下では $|C_{ML}| = |C_{CL}|$ とし，制約数を変えて実験した．

データ対間の距離尺度としては，文献 16) 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重み行列 $\mathbf{W}$ を定義した．

提案法と 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．提案法に対しては完全グラフを構築したが，完全グラフを用いて PCP を実行したところ非常に精度が悪かった．このため，文献 15) 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，また文献 15) に従って近傍数 $m = 10$ とした．

提案法と 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．次元数も結果に影響をおよぼすが，次元数 l = クラスタ数 k とした．

SCREEN では線形写像を求める際に 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）に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 16) ではまず前処理として主成分分析（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 16) に従い，寄与率の観

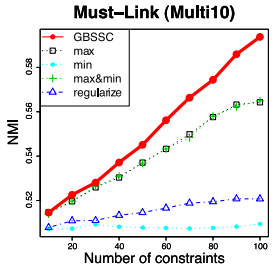

図 5　重み変更の比較
Fig. 5　Weight modifications.

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 C_{ML} と 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．さらに，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，その平均を求めた．

4.2　グラフ表現に基づく射影の評価

提案法は must-link を反映させてデータ間の重み（類似度）を定義し，cannot-link を反映して射影表現を求めている．以下ではそれぞれに対する評価を述べる．本節の図では縦軸は精度（NMI）を示す．

4.2.1　must-link に基づく縮約の効果

must-link を反映させるため，提案法では縮約を適用し，縮約後のグラフに対して式 (9) により重みを定義した．式 (10) のように制約を正則化項として扱うのではなく，3.3 節で述べたグラフの重みを変更する際の他の定義方法として，縮約を用いずに，データ集合に対する重み付きグラフにおいて，i) must-link C_{ML} に含まれる各ペアの重みを類似度最大値（$w_{ij} = 1$）に設定する，ii) cannot-link C_{CL} に含まれる各ペアの重みを類似度最小値（$w_{ij} = 0$）に設定する，というアプローチが考えられる．上記により重みを修正し，正規化グラフラプラシアンを用いてスペクトルクラスタリングを行った場合と提案法を比較した．

Multi10 に対する結果を**図 5** に示す（横軸は制約数）．図中の凡例で max は上記の i)，min は ii) に対応する．max&min は i)，ii) を併用した場合に対応する．GBSSC は提案法における縮約を用いた場合である[⋆1]．なお，regularize は縮約を用いずに式 (11) を用いて正則化した場合に対応する[⋆2]．図 5 に示すように提案法は上記 i)，ii) などよりも精度が上回っており，must-link を縮約し，式 (9) により縮約後のグラフに対する重み（類似度）を設定する提案法は有効であると考えられる．

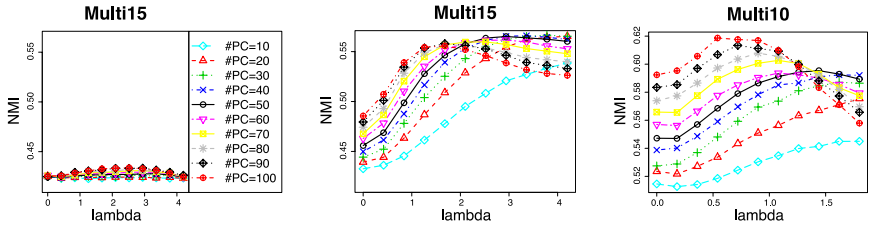

図 6　式 (10) での λ の影響（左図：式 (11) を使用，中図・右図：式 (12) を使用）
Fig. 6　Influence of λ in Eq. (10).

上記の結果に対する要因として，上記 i) を用いた場合には must-link でつながれたデータ対の重みだけが影響を受けるため，must-link でつながれたデータだけからなる孤立した小さなクラスタが生成されやすくなると考えられる．他方，提案法では must-link でつながれたデータ対だけではなく，この制約が指定されたデータに隣接するデータとの重み（類似度）も更新されて制約の影響を受けるため，この問題を回避できると考えられる．

4.2.2　cannot-link に基づく正則化の効果

提案法では cannot-link の影響（式 (10) の第 2 項）を λ で重み付けしている．cannot-link に基づく正則化の効果を調べるため，must-link に関するグラフ縮約を行った後に，行列 **S** を式 (11)，(12) で定義した場合に λ の値を変えて実験し，その影響を調べた．結果を**図 6** に示す．図 6 の左図が式 (11)，中図と右図が式 (12) に対応し，横軸は λ である．図中の凡例で#PC は制約数を示す．

図 6 より，従来のように式 (11) を用いた場合（左図）は制約数や λ による精度の違いはほとんどみられないが，提案法（縮約と式 (12) の行列 **S**）では λ の値に応じた精度の向上がみられた（図 6 の中，右図）．また，提案法で Multi10 と Multi15 を比較するとクラスタ数の増加にともないより大きな λ で精度向上が実現されている．この理由として，cannot-link

⋆1 must-link に対する縮約の効果を評価するため，cannot-link は用いていない（$\lambda = 0$ とした）．
⋆2 図 6 に示すように式 (11) を用いた場合は λ の影響が少ないため，$\lambda = 0.9$ の場合を示す．

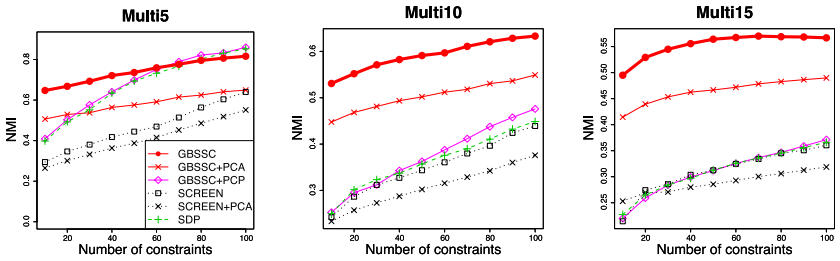

図 7　20 Newsgroup に対する結果（NMI）
Fig. 7　Results on 20 Newsgroup datasets (NMI).

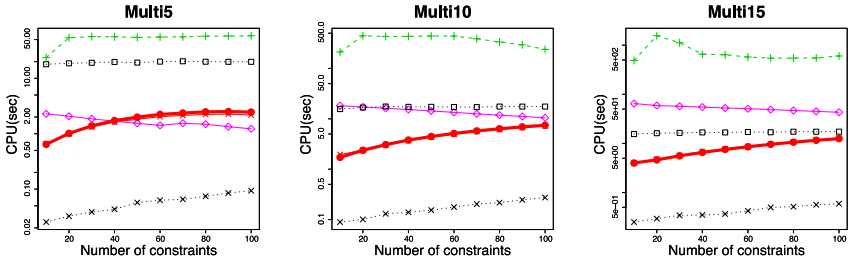

図 8　20 Newsgroup に対する結果（CPU 時間）
Fig. 8　Results on 20 Newsgroup datasets (CPU time).

はクラスタ間での関係（クラスタの組）を表現するが可能な組合せはクラスタ数に依存し，個々の制約からの影響は組合せ数が多いほど相対的に小さくなる可能性が考えられる．以下ではクラスタの組合せ数を考慮して正則化項を活用するために $\lambda = \lambda_0 \cdot {}_kC_2$ と設定し（k はクラスタ数），予備実験から $\lambda_0 = 0.02$ として実験した．

4.3　実データに対する評価

4.1.1 項でのデータセットに対する結果を示す．本節の図で横軸は制約数に対応する．縦軸は式 (13) で定義した NMI，あるいは試行 1 回に要する CPU 時間（秒）の平均値に対応する．図中の凡例で+PCA は主成分分析を用いて低次元表現を生成し，その表現に対して各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 に対して縮約を行い，cannot-link に対しては文献 15) に従って m-近傍グラフを構築して PCP を適用した場合を示す．

4.3.1　20 ニュースグループ

表 1 に示した 20NG に対する結果（精度，実行速度）を**図 7**，**図 8** に示す．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のクラスタリングに対応するが，図 7 より，部分空間の次元数 l=クラスタ数 k とした場合[⋆1]に提案法（GBSSC，赤太線）は他手法を上回る性能 (精度) を示した．図 7 で Multi5 に対しては制約数が増加するにつれて PCP（緑破線）は GBSSC に近づき，制約数が 70 を超えるとほぼ同程度の性能を示したが，図 8 に示すように提案法は 2 桁（100 倍）程度高速であった．

提案法を用いて must-link を縮約して cannot-link に PCP を適用した場合（GBSSC+PCP，桃線）は，精度は PCP とほぼ同程度あったが，must-link の縮約を行うことにより PCP と比較して 1 桁（10 倍）程度高速であった．しかし PCP はデータ数に応じて飛躍的に計算時間が増加するため，データ数が多い Multi15 では提案法が GBSSC+PCP よりもさらに 1 桁（10 倍）程度高速であり，また精度も大幅に上回った．

主成分分析の使用は SCREEN（黒点線）の高速化には効果があったが，提案法に対しては効果がなく（GBSSC+PCA，赤線），精度は逆に悪化した．このため，提案法に対しては前処理として主成分分析を用いて低次元表現を生成することは不要であるといえる．

4.3.2　TREC データセット

表 2 に示した TREC データセットに対する結果（精度，実行速度）を**図 9**，**図 10** に示す[⋆2]．

20NG に対する結果と同様，図 9 より精度に関しては提案法（GBSSC）は SCREEN を大きく上回る性能を示した．他方，PCP との比較では，tr11，tr12，tr41，tr45 に対しては制約数が少ない場合には提案法は PCP を上回ったが，tr23，tr31 に対しては下回った．つねに PCP を上回るわけではないが，少量の制約を用いて性能向上を実現するという半教師ありクラスタリングの観点からは提案法は効果的であると考えられる．また，図 10 に示すように実行速度に関しては 20NG とほぼ同様な結果となり，提案法は PCP に比べて 2 桁（100 倍）以上高速であった．提案法と PCP を組合せた GBSSC+PCP は，20NG の場合とほぼ同様な傾向を示した．

⋆1 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．

⋆2 SCREEN では PCA を用いないと非常に時間がかかってしまうため，PCA を用いる場合のみ評価した．

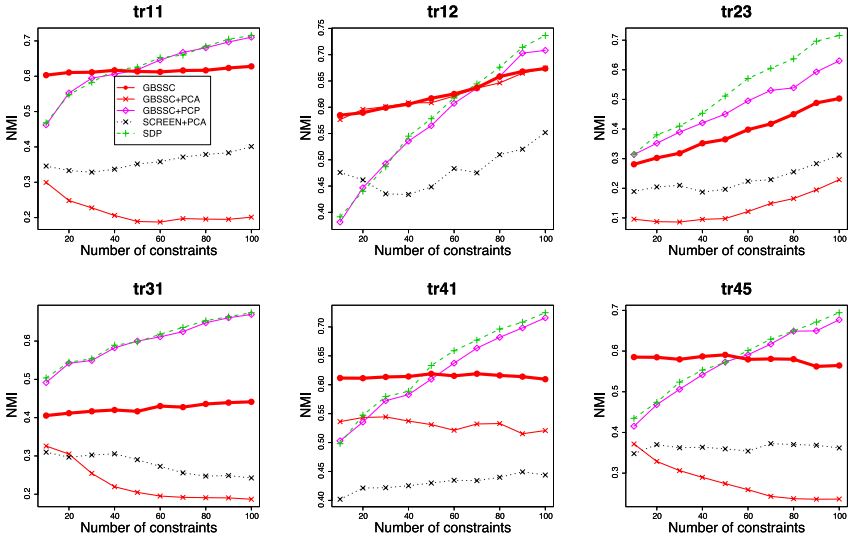

図 9　TREC データセットに対する結果（NMI）
Fig. 9　Results on TREC datasets (NMI).

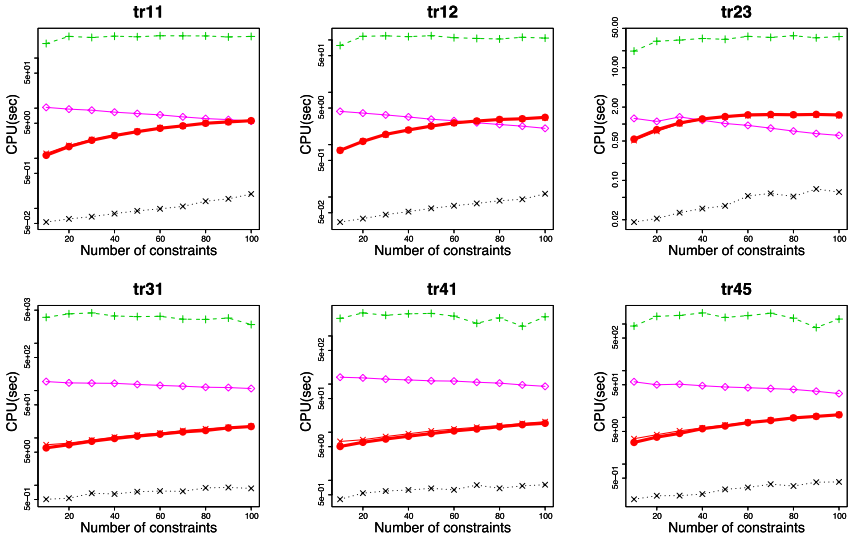

図 10　TREC データセットに対する結果（CPU 時間）
Fig. 10　Results on TREC datasets (CPU time).

4.4 考　　察

4.3 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および実行速度の観点から提案法の有効性を確認した．SCREEN [16] に対しては，提案法は精度および実行速度の観点で大きく上回る性能を示した．PCP [15] 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）程度高速であった．特に，少量の制約を用いる場合には精度の観点からも PCP を上回る性能を示した．上記より，提案法は半教師ありクラスタリング手法として有効であると考えられる．

2.2 節で述べたように，SCREEN [16] は制約に基づいて分散が最大化されるような部分空間にデータ全体を射影し，射影表現に対してクラスタリングを行う．他方，提案法は縮約とグラフスペクトルに基づく次元縮約 [17] を行い，制約に基づいて式 (10) が最小化される部分空間にデータ全体を射影し，射影表現に対してクラスタリングを行う．must-link に対する縮約を行うことによりデータ数を削減するだけではなく，データ対間の関係（類似度と制約）に基づいたグラフ表現を構築する．さらに，式 (12) で定義した行列 $\mathbf{S}$ を用いて cannot-link をその近傍のデータ対にも拡張して正則化を行う．

PCP はデータ集合に対して制約を反映する空間でのデータ間の類似度を半正定値計画問題を解いて求め，カーネル kmeans を用いてクラスタリングを行うが，半正定値計画問題を解くため計算コストの高い手法であり，大規模なデータには適用しにくいという課題がある．実験結果からも，クラスタ割当ての精度が高い反面，多大な計算コストを要することが分かる．実行速度と精度とのトレードオフを考慮すると，提案法は高次元データに対する半教師ありクラスタリング手法として有効であると考えられる．

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，制約を反映させ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を用いることによりそれぞれの制約を反映させ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．提案法を 20 ニュースグループや TREC データセットなどの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有効性を確認した．今後は画像などの他の実データに対しても評価を行いさらなる改良を行う予定である．特に，cannot-link に基づく正則化の拡張に取り組んでいきたい．

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．最後に，有益なご指摘を賜りました査読者の方々に深く謝意を表します．

参 考 文 献

1) Basu, S., Bilenko, M. and Mooney, R.J.: A Probabilistic Framework for Semi-Supervised Clustering, *Proc. KDD'04*, pp.59–68 (2004).
2) Belkin, M. and Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural Computation*, Vol.15, pp.1373–1396 (2002).
3) Bie, T.D., Suykens, J. and Moor, B.D.: Learning from General Label Constraints, *LNCS* 3138, pp.671–679 (2004).
4) Blum, A. and Mitchell, T.: Combining Labeled and Unlabeled Data with To-Training, *Proc. COLT'98*, pp.92–100 (1998).
5) Chapelle, O., Schölkopf, B. and Zien, A. (Eds.): *Semi-Supervised Learning*, MIT Press (2006).
6) Chung, F.: *Spectral Graph Theory*, American Mathematical Society (1997).
7) Dhillon, J. and Modha, D.: Concept decompositions for lage sparse text data using clustering, *Machine Learning*, Vol.42, pp.143–175 (2001).
8) Dhillon, J., Mallela, S. and Modha, D.: Information-theoretic co-clustering, *Proc. KDD'03*, pp.89–98 (2003).
9) Diestel, R.: *Graph Theory*, Springer (2006).
10) Elghazel, H., Yoshida, T., Deslandres, V., Hacid, M. and Dussauchoy, A.: A new greedy algorithm for improving b-coloring clustering, *Proc. GbR'07*, pp.228–239 (2007).
11) Goldberg, A.B., Zhu, X. and Wright, S.: Dissimilarity in Graph-Based Semi-Supervised Classification, *Proc. AISTAT'07*, pp.155–162 (2007).
12) Guënoche, A., Hansen, P. and Jaumard, B.: Efficient algorithms for divisive hierarchical clustering with the diameter criterion, *J. Classification*, Vol.8, pp.5–30 (1991).
13) Hartigan, J. and Wong, M.: Algorithm AS136: A k-means clustering algorithm, *Journal of Applied Statistics*, Vol.28, pp.100–108 (1979).
14) Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., Luca, E.W.D. and Albayrak, S.: Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization, *Proc. SDM'10*, pp.559–570 (2010).
15) Li, Z., Liu, J. and Tang, X.: Pairwise constraint propagation by semidefinite programming for semi-supervised classification, *Proc. ICML-08*, pp.576–583 (2008).
16) Tang, W., Xiong, H., Zhong, S. and Wu, J.: Enhancing Semi-Supervised Clustering: A Feature Projection Perspective, *Proc. KDD'07*, pp.707–716 (2007).
17) von Luxburg, U.: A Tutorial on Spectral Clustering, *Statistics and Computing*, Vol.17, No.4, pp.395–416 (2007).
18) Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S.: Constrained K-means Clustering with Background Knowledge, *Proc. ICML'01*, pp.577–584 (2001).
19) Wang, J., Kumar, S. and Chang, S.-F.: Sequential Projection Learning for Hashing with Compact Codes, *Proc. ICML'10* (2010).
20) Xing, E.P., Ng, A.Y., Jordan, M.I. and Russell, S.: Distance metric learning, with application to clustering with side-information, *Proc. NIPS15*, pp.505–512 (2003).

(平成 22 年 8 月 31 日受付)
(平成 22 年 10 月 5 日再受付)
(平成 22 年 10 月 21 日採録)

吉田　哲也（正会員）

1968 年生．1991 年東京大学工学部航空工学科卒業．1997 年東京大学大学院博士課程修了．工学博士．同年大阪大学大学院基礎工学研究科助手．2001 年大阪大学産業科学研究所助手．2004 年北海道大学大学院情報科学研究科助教授．現在，同大学准教授．主に機械学習，知識獲得，データマイニング等の研究に興味を持つ．人工知能学会会員．

岡谷　一宏

1984 年生．2008 年北海道大学工学部電子工学科卒業．2010 年北海道大学大学院情報科学研究科修士課程修了．現在，日立情報制御ソリューションズ勤務．主に機械学習や通信制御等に興味を持つ．