Vol.2009-MPS-76 No.30
Vol.2009-BIO-19 No.30
2009/12/18

相互情報量に基づくクラスタリングに対するグラフモデルとその評価

吉 田 哲 也[†1]

本稿では，相互情報量に基づくクラスタリング問題に対するグラフモデルを提案する．相互情報量から導出される定常分布に着想を得たデータ間の類似度関数を定義してデータ集合を辺重み付きグラフとして表現することにより，データが一様分布する場合にはハードクラスタリング問題が提案するグラフモデルにおける組合せ最適化問題に近似できることを示す．提案するグラフモデルを文書クラスタリングでのベンチマークデータである 20 Newsgroup のデータに対して評価し，他手法との比較を通じて提案手法の妥当性と有効性を確認した．

A Graph Model for mutual information based clustering and its evaluation

Tetsuya Yoshida [†1]

We propose a graph model for data clustering based on mutual information. Based on the stationary distribution induced from the problem setting, we propose a similarity function among data objects, and represent the entire objects as an edge-weighted graph. We show that, in hard assignment, the problem can be approximated as a combinatorial problem over the proposed graph when data is uniformly distributed. The proposed approach is evaluated on the text clustering problem over the 20 Newsgroup benchmark data. The results are encouraging and indicate the effectiveness of our approach.

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

1. はじめに

本稿では，文献[10] で提案された相互情報量に基づくクラスタリングの枠組み[10] について考察する．この枠組みでは，データクラスタリングは相互情報量に基づく制約付最適化問題として定式化されるが，相互情報量の非線形性や目的関数の非凸性などにより大域的最適解を得ることが困難なため，様々な近似解法が提案されてきた[7),9),10)]

本稿では，相互情報量に基づくクラスタリング問題に対するグラフモデルを提案し，データが一様分布する場合にはハードクラスタリング問題が提案するグラフモデルにおける組合せ最適化問題に近似できることを示す．まず，相互情報量から導出される定常分布に着想を得たデータ間の類似度関数を定義し，この類似度関数に基づいてデータ集合全体を辺重み付きグラフとして表現することを提案する．各データをグラフの頂点に対応させ，頂点対を類似度関数から得られる類似度を重みとする辺で連結することにより，データ集合を辺重み付きグラフとして表現する．次に，相互情報量に基づくクラスタリング問題が，提案するグラフモデルによりデータ集合を表現したグラフ上の組合せ最適化問題として定式化できることを示す．

データ集合を本稿で提案するグラフモデルに基づいて辺重み付きグラフとして表現することにより，グラフ構造に基づく様々なアルゴリズムを用いてクラスタリングを行うことが可能になる．たとえば，グラフの最小カット問題に対応するスペクトルクラスタリング法[11)] を用いて文献[10] での制約付最適化問題を解くことが可能となる．提案手法を文書クラスタリングでのベンチマークデータである 20 Newsgroup のデータに対して評価し，他手法との比較を通じて提案手法の妥当性と有効性を確認した．特に，クラスタ数が多く困難な問題に対する提案手法の有効性を確認した．

2. 問題設定

2.1 準　　備

以下では，$\boldsymbol{X}$ で (与えられた) データ集合を表現し，$|\boldsymbol{X}|$ で集合の大きさ (要素数) を表現する．$\mathcal{X}$ 上の確率変数 X に対する確率分布 $p_1(x)$ と $p_2(x)$ を考える．

定義 1. $\mathcal{X}$ 上の確率変数 X に対する確率分布 $p_1(x)$ と $p_2(x)$ に対する *Kullback-Leibler (KL)* 情報量は以下で定義される[1)]．

$$D_{KL}[p_1(x)||p_2(x)] = \sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)} \tag{1}$$

Vol.2009-MPS-76 No.30
Vol.2009-BIO-19 No.30
2009/12/18

$\mathcal{X}$，$\mathcal{Y}$ 上の確率変数 X，Y に対し，結合確率分布を $p(x,y)$ と表記し，$p(x)$ と $p(y)$ をその周辺確率分布とする．また，条件付確率分布を $p(y|x)$ と表記する．

定義 2. *2* つの確率変数 X と Y の間の相互情報量 $I(X;Y)$ は以下で定義される．

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

$$= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \tag{3}$$

$$= D_{KL}[p(x,y)||p(x)p(y)] \tag{4}$$

2.2 情報ボトルネック法

相互情報量に基づくクラスタリングの枠組みは文献[10] で情報ボトルネック法として提案された．この手法は文献[5] の手法と同様，データ集合$\boldsymbol{X}$の確率分布に基づいてクラスタリングを行うが，与えられたデータ集合に対する関連変数 Y を導入し，Y に対する情報を多く持つようなクラスタの集合$\boldsymbol{T}$を求める問題と捉える枠組みである．

この枠組みにおいては，確率変数 T は X のみに依存して Y には無関係である，ということが以下のマルコフ関係として定式化される．

定義 3. 確率変数 X, Y, T に対して以下のマルコフ関係が成り立つ．

$$T \leftrightarrow X \leftrightarrow Y \tag{5}$$

一例として，文書集合$\boldsymbol{X}$=$\{x_1,\ldots,x_n\}$ が与えられ，各文書 x_i は文書中の単語の頻度ベクトルとして表現される場合を考える．この場合，文書集合$\boldsymbol{X}$を表現するのに使用された単語の集合が$\boldsymbol{Y}$=$\{y_1,\ldots,y_m\}$ に対応する．$p(x,y)$ は文書 x と単語 y の同時確率に対応し，たとえば文書 x と単語 y の共起回数に基づいて推定される．クラスタリングの目的は，各クラスタ t が単語の予測に有効であるような$\boldsymbol{X}$に対するクラスタ集合$\boldsymbol{T}$=$\{t_1,\ldots,t_k\}$ を見つけることである．

定義 3 に基づき，文献[10] は相互情報量に基づくクラスタリングを制約付最適化問題として定式化した．

問題 1. 以下の目的関数 $\mathcal{L}$ を最小化する条件付確率 $p(t|x)$ を求めよ．

$$\mathcal{L} = I(X;T) - \beta I(T;Y) \tag{6}$$

$I(X;T)$ と $I(T;Y)$ はそれぞれ X と T，T と Y の相互情報量であり，β はハイパーパラメータである．

問題 1 での枠組みを図 1 に示す． 直観的には，データ集合$\boldsymbol{X}$をクラスタ集合$\boldsymbol{T}$に圧縮して表現し，圧縮した表現$\boldsymbol{T}$が$\boldsymbol{Y}$ について情報を多く持つ（予測できる）という問題設定を，$\boldsymbol{X}$, $\boldsymbol{T}$, $\boldsymbol{Y}$に対応する確率変数 X, T, Y を考え，圧縮の程度を相互情報量 $I(X;T)$ で表現し，また予測の程度を相互情報量 $I(T;Y)$ で表現することにより，両者を相互情報量に基づくクラスタリング問題として定式化している．

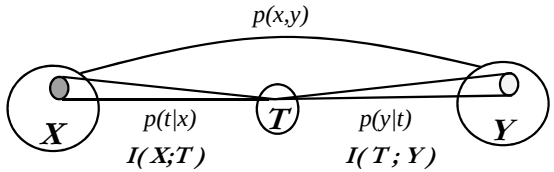

図 1 相互情報量に基づくクラスタリングの枠組み[10)]
Fig. 1 Data clustering based on mutual information[10)]

問題 1 に対する最適解は以下に示す式を満たす必要がある[10)]．

定理 1. $p(x,y)$ と β が与えられ，マルコフ関係 *(5)* が成立する場合には，$p(t|x)$ は以下の式を満たす場合に限り $\mathcal{L}$ の定常分布となる．

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} exp(-\beta D_{KL}[p(y|x)||p(y|t)]) \tag{7}$$

$$Z(x,\beta) = \sum_t p(t) exp(-\beta D_{KL}[p(y|x)||p(y|t)]) \tag{8}$$

2.3 従来の近似解法

定理 1 は問題 1 に対する解 $p(t|x)$ が前節の問題設定における定常分布であることを示すが，式 (7) における左辺の $p(t|x)$ は同時に右辺にも（非線形に）影響を及ぼす．更に，式 (6) での目的関数 $\mathcal{L}$ は $p(t|x)$, $p(t)$, $p(y|t)$ に対して同時には凸ではない．このため，問題 1 に対する大域的最適解を求めることは非常に困難な問題となる．

上記の問題に対し，これまで近似解を求めるいくつかの解法が提案されてきた[7),9),10)]．その中で，sIB 法と呼ばれる手法が計算量や得られるクラスタの質の観点から他の手法よりも良いことが報告されている[9)]．この手法は，問題 1 に双対な最大化問題をデータの逐次再割り当てによって近似的に 解く手法であり，各データを単一のクラスタに割り当てるという意味でハードクラスタリングを行う．

Vol.2009-MPS-76 No.30
Vol.2009-BIO-19 No.30
2009/12/18

3. 提案するグラフモデル

3.1 準　　備

頂点集合$\boldsymbol{V}$と辺集合$\boldsymbol{E}$ $\in \boldsymbol{V} \times \boldsymbol{V}$ から構成されるグラフを$G(\boldsymbol{V}, \boldsymbol{E})$ と表記する．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，重みの集合を$\boldsymbol{W}$とする．$|\boldsymbol{V}| = n$ の場合，$\boldsymbol{W}$の重みは $n \times n$ 行列 $\mathbf{W}$ で表現することができ，$\mathbf{W}$ 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．なお，辺がない頂点対間での重みは 0 とする．

3.2 データ間の類似度関数

本稿では，定理 1 と式 (7) に基づき，データ x とクラスタ t の間の KL 情報量 $D_{KL}[p(y|x)||p(y|t)]$ が 2 節での枠組みにおける非類似度を表現すると捉える．更に，この非類似度を $\mathcal{X} \times \mathcal{T}$ から $\mathcal{X} \times \mathcal{X}$ に拡張して，KL 情報量を問題 1 におけるデータ間での非類似度を表現する関数と捉える．

上記の非類似度関数に基づき，2 節の枠組みにおけるデータ間での類似度関数として以下を提案する．

定義 4. *s:* $\mathcal{X} \times \mathcal{X} \to \mathcal{R}^+$ は以下で定義されるデータ間の類似度関数である．

$$s(x_i, x_j) = p(x_j) exp(-\beta D_{KL}[p(y|x_i)||p(y|x_j)]) \tag{9}$$

β は問題 *1* でのハイパーパラメータである．

3.3 データグラフ

式 (9) で定義した類似度関数は，データ集合$\boldsymbol{X}$での任意のデータ対 (x_i, x_j) の間の関係を類似度として表現する．データ対の関係はグラフとして表現できるため，2 節の問題設定においてデータ集合$\boldsymbol{X}$を式 (9) で計算される類似度を重みとする辺重み付きグラフとして表現するグラフモデルを提案する．

定義 5. データ集合$\boldsymbol{X}$に対する辺重み付きグラフ $G(\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{W})$ を以下で定義する．

$$\boldsymbol{V} = \boldsymbol{X} \tag{10}$$

$$w_{ij} = \begin{cases} s(x_i, x_j) & x_i \neq x_j \\ 0 & otherwise \end{cases} \tag{11}$$

$$\boldsymbol{E} = \{(x_i, x_j) | w_{ij} > 0\} \tag{12}$$

式 (10) より各データと頂点は 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 4 より全ての重みは非負（$w_{ij} \geq 0, \forall x_i, x_j \in \boldsymbol{X}$）である．本稿では，上記で定義したグラフを**データグラフ**と呼ぶ．以下では，データグラフ G は連結グラフであると仮定する[⋆1]．

命題 2. $\frac{w_{ij}}{\sum_j w_{ij}}$ はデータグラフにおいて頂点 x_i と x_j の間の条件付確率である．

証明　略

命題 2 より，データグラフにおける条件付確率を以下で定義する．

$$p(x_j|x_i) = \frac{w_{ij}}{\sum_j w_{ij}} \tag{13}$$

式 (13) の条件付確率はデータグラフ上でのデータ x_i からデータ x_j への遷移確率と解釈できる．

命題 3. 式 *(13)* の条件付確率は$\boldsymbol{T} = \boldsymbol{X}$の場合における定理 *1* の定常分布である．

証明　略

$\boldsymbol{T} = \boldsymbol{X}$とすることはデータ集合に対する圧縮 (分割) が全く行われないことに対応する．この場合には，提案した類似度関数（式 (9)）から導出されるデータグラフにおいて式 (13) で定義される $p(x_j|x_i)$ が定理 1 での条件を満たすことを命題 3 は示している．

3.4 データグラフに基づくアプローチ

本稿では，ハードクラスタリングにおける問題 1 はデータが一様分布する場合には提案するグラフモデルにおける以下の問題に近似できることを示す．

問題 2. クラスタ数 k が指定された場合に，データグラフ G において以下の目的関数 J を最小化する互いに素な辺の集合族 $\{\boldsymbol{E}_1, \ldots, \boldsymbol{E}_k\}$ を求めよ．

$$J = \sum_{t=1}^{k} \sum_{w_{ij} \in \boldsymbol{E}_t} w_{ij} \tag{14}$$

ただし，G から $\{\boldsymbol{E}_1, \ldots, \boldsymbol{E}_k\}$ を削除すると G は k 個の連結要素に分割されるものとする．

与えられたデータにおいて，データ集合 $\boldsymbol{X}$ とその関連変数 $\boldsymbol{Y}$ に対応する確率変数 X と Y との相互情報量 $I(X;Y)$ はある定数となる．このため，任意の β に対して以下の問題は問題 1 と同値な問題である[8)]．

問題 3. 以下の目的関数 F_{IB} を最小化する $p(t|x)$ を求めよ．

$$F_{IB} = \sum_x \sum_t p(x)p(t|x)(-\log Z(x, \beta)) \tag{15}$$

証明[8)]　略

⋆1 各連結要素ごとにデータグラフを考えれば一般性を失わない．

Vol.2009-MPS-76 No.30
Vol.2009-BIO-19 No.30
2009/12/18

データグラフ G においては，式 (15) の目的関数は以下で表現される．

$$F_G = \sum_{x_i} \sum_{x_j} p(x_i)p(x_j|x_i)(-\log Z(x_i, \beta)) \tag{16}$$

頂点 x_i から出る辺の重みの総和を d_i と定義する [⋆1]．

$$d_i = \sum_{x_j} w_{ij}, \quad \forall x_i \in \boldsymbol{X} \tag{17}$$

本稿での主要結果を示すために以下を仮定する．

仮定 1. データは一様分布し，$p(x)$ はある定数 $c > 0$ である．以下では，この仮定を一様分布と呼ぶ．

命題 4. 一様分布の下では，データ集合$\boldsymbol{X}$ に対するデータグラフ G において F_G はある定数である．

証明 略

3.4.1 データ圧縮とカット

データ集合 $\boldsymbol{X}$ を$\boldsymbol{X}= S \sqcup \bar{S}$ [⋆2] という 2 つの部分集合へ分割することを考える[11)]．データグラフ G において部分集合 S と $\bar{S}$ の間を連結する辺を削除することにより，G は対応する誘導部分グラフ G_S と $G_{\bar{S}}$ に分割され，分割後のグラフは (非連結な) グラフ $\hat{G} = \{G_S, G_{\bar{S}}\}$ となる[3)]．

定義 6. 辺の削除による分割に対して $cut(S, \bar{S})$，$cut(\bar{S}, S)$ を以下で定義する．

$$cut(S, \bar{S}) = \sum_{x_i \in S} \sum_{x_j \in \bar{S}} w_{ij} \tag{18}$$

$$cut(\bar{S}, S) = \sum_{x_i \in \bar{S}} \sum_{x_j \in S} w_{ij} \tag{19}$$

命題 5. 各部分集合の要素数が *1* より大きい任意のデータグラフ G の分割に対し，$\frac{w_{ij}}{\sum_{j \in G_S} w_{ij}}$ は各誘導部分グラフ G_S における条件付確率である．

証明 略

データ集合の分割$\boldsymbol{X}= S \sqcup \bar{S}$ において，データ集合 $\boldsymbol{X}$ の各要素 x_i に対して x_i を含む部分集合を S_i と表記し，含まない部分集合を $\bar{S}_i$ と表記する．

式 (17) と同様に以下を定義する．

$$d_{S_i} = \sum_{x_j \in S} w_{ij} \tag{20}$$

$$d_{\bar{S}_i} = \sum_{x_j \in \bar{S}} w_{ij} \tag{21}$$

各データ $x_i \in \boldsymbol{X}$ に対し，式 (17)，(20)，(21) の間に以下の関係が成り立つ．

$$d_i = d_{S_i} + d_{\bar{S}_i} \tag{22}$$

命題 5 に基づき，分割後のグラフ $\hat{G}$ における条件付確率を以下で定義する．

$$\forall x_i \in S, \quad \hat{p}(x_j|x_i) = \begin{cases} \frac{w_{ij}}{d_{S_i}} & x_j \in S \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

$$\forall x_i \in \bar{S}, \quad \hat{p}(x_j|x_i) = \begin{cases} \frac{w_{ij}}{d_{S_i}} & x_j \in \bar{S} \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

式 (16) と同様に F_{G_S} と $F_{G_{\bar{S}}}$ を定義すると，一様分布の下では以下が成り立つ．

$$F_{\hat{G}} = F_{G_S} + F_{G_{\bar{S}}} \tag{25}$$

ただし，各データ x_i は S_i のみに割り当てられるために $p(S_i|x_i) = 1$，$p(\bar{S}_i|x_i) = 0$, $\forall x_i \in \boldsymbol{X}$ となる[⋆3]．このため，式 (23)，(24) で定義される $\hat{p}(x_j|x_i)$ は式 (7) を満たさず，問題 3 に対する最適解から乖離することになる[⋆4]．

問題 3 を解くために分割に伴う最適解からの乖離を最小化することを考える．命題 4 よりデータ集合$\boldsymbol{X}$に対するデータグラフ G において F_G は定数であるため，分割に伴う最適解からの乖離 $F_{\hat{G}} - F_G$ の最小化は以下と同値な問題となる．

問題 4. データグラフ G において，以下を最小化する分割$\boldsymbol{X}= S \sqcup \bar{S}$ を求めよ．

$$F_{\hat{G}} = F_{G_S} + F_{G_{\bar{S}}} \tag{26}$$

ここで $\hat{G} = \{G_S, G_{\bar{S}}\}$ である．

3.4.2 主 要 結 果

本稿での主要な結果を示す．まず，以下の問題を定義する．

問題 5. データ集合$\boldsymbol{X}$に対するデータグラフ G において，以下の目的関数 J を最小化する互いに素な辺の集合族 $\{\boldsymbol{E}_1, \boldsymbol{E}_2\}$ を求めよ．

⋆1 $\sum_{x_j}$ は全ての$\boldsymbol{X}$に渡る和であり，$\sum_j$ に対応する．
⋆2 部分集合 $\bar{S}$ は部分集合 S の補集合に対応する．

⋆3 S と $\bar{S}$ は 2 つのクラスタに対応し，この 2 つのクラスタへのハードクラスタリングに対応する
⋆4 一般にハードクラスタリングでは最適解から乖離する．

Vol.2009-MPS-76 No.30
Vol.2009-BIO-19 No.30
2009/12/18

$$J = \sum_{t=1}^{2} \sum_{w_{ij} \in \boldsymbol{E}_t} w_{ij} \tag{27}$$

ただし，G から $\{\boldsymbol{E}_1, \boldsymbol{E}_2\}$ を削除すると G は 2 個の連結要素に分割されるものとする．

主張 6. ハードクラスタリングにおいては，一様分布の下では問題 *1* は問題 *5* に近似できる．

証明 略

主張 6 は以下に拡張できる．

主張 7. ハードクラスタリングにおいては，一様分布の下では問題 *1* は問題 *2* に近似できる．

3.5 データグラフに基づくクラスタリング

前節より，提案するグラフモデルに基づいてデータ集合$\boldsymbol{X}$をデータグラフとして表現することにより，2 節での相互情報量に基づくクラスタリング問題をデータグラフにおける組合せ最適化問題 (問題 2) の観点からアプローチすることが可能となる．たとえば，この問題を効率的に解く 様々なアルゴリズムの利用が考えられる．

ただし，カットに基づくクラスタリングの定式化においては非常に小さなクラスタが生成されてクラスタ集合のサイズに偏りが生じるという問題がある[11]．データ集合をいくつかのクラスタに分割するというクラスタリングの観点からは，クラスタの偏り (極端にサイズの小さなクラスタの生成など) は望ましいことではないと考えられる．このため，データグラフに基づいて相互情報量に基づくクラスタリング問題を解く際には，目的関数の最小化に加えてクラスタ相互のバランスを考慮することが重要となると考えられる．

4. 評　　価

4.1 文書クラスタリングへの適用

先行研究[8] に基づき，提案したグラフモデルを文書クラスタリングに適用して評価した．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り，各文書 x は文書処理で標準的な単語の頻度に基づくベクトル空間モデルで表現されると仮定する．2.2 節での例と同様，$\boldsymbol{X}$を表現する全単語集合が$\boldsymbol{Y}$=$\{y_1, \ldots, y_m\}$ に対応し，$p(x, y)$ は文書 x と単語 y の同時確率に対応する．一般に文書に含まれる単語数は膨大である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．本稿の手法は分割的クラスタリングに対応するためクラスタ数 k=$|\boldsymbol{T}|$ は与えられると仮定する．

評価対象として，文書クラスタリングのベンチマークである 20 ニュースグループ (以下，20NG) ⋆1を使用した． 先行研究[8] を参考に，本稿では 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．各データセットに対して porter stemmer ⋆2 を用いて stemming を行い，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．

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

4.2 実 験 設 定

4.2.1 手　　法

各データセットに対して 3.3 節のデータグラフを作成し，問題 2 に対応する解を求めてクラスタリングを行った．3.5 節で述べたように，データ集合をクラスタリングする際には処理の目的を反映して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と考えられる．本稿では，この点を考慮した手法としてスペクトルクラスタリングを用いた[11]．提案するデータグラフでは頂点対 (x_i, x_j) に対してそれぞれ w_{ij},w_{ji} を持つ辺が定義されるが，分割の際にはその両方を削除する必要がある．このため，以下の実験では式 (11) での重みに基づいて対称行列 $(\boldsymbol{W})_{ij}$=$(w_{ij} + w_{ji})/2$ を作成した．

クラスタ相互のバランスを考慮するために，式 (17) で定義される d_i を対角要素とする対角行列 $\boldsymbol{D}$ を用いて正規化した以下の 2 つが代表的な手法として提案されている[11]．

$$\boldsymbol{L_{rw}} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W} \tag{28}$$

$$\boldsymbol{L_{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}} \tag{29}$$

⋆1 http://people.csail.mit.edu/ jrennie/20Newsgroups/．本稿では 20news-18828 を使用した．
⋆2 http://www.tartarus.org/ martin/PorterStemmer

Vol.2009-MPS-76 No.30
Vol.2009-BIO-19 No.30
2009/12/18

それぞれを用いて表現 $\boldsymbol{H_{rw}}$,$\boldsymbol{H_{sym}}$ を生成し，これらに対して kmeans 法を用いてクラスタリングを行った．

比較手法として，文献[9),10)] で提案された iIB 法, sIB 法，および高次元スパースデータに対する標準手法である skmeans 法[2)] での実験を行った．iIB 法は式 (7) の定常分布を交互射影により求める手法であり，sIB 法は問題 1 と双対な問題をデータの逐次再割り当てにより求める手法である．

なお，テキスト処理でしばしば遭遇するゼロ頻度問題のために KL 情報量は数値的に不安定となる恐れがある．このため，Ristad 法[6)] でのスムージングを用いて各データセットでの文書 x と単語 y の同時確率 $p(x,y)$ を推定した．

4.2.2 評 価 尺 度

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各手法が割り当てるクラスタに基づいて以下で述べる正規化相互情報量 (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．

$$NMI = \frac{I(\hat{T};T)}{(H(\hat{T})+H(T))/2} \quad (\in [0,1]) \tag{30}$$

$H(T)$ 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．

なお，生成されたクラスタのまとまり具合に対応する純度 (purity)[4)] についても評価し，提案手法と他手法との比較を行ったが，紙面の都合から本稿では報告を割愛する．

4.2.3 パラメータ

問題 1 で述べたように，相互情報量に基づくクラスタリングの枠組みにおける主要なパラメータはハイパーパラメータ β であり，式 (9) でも用いられる．sIB 法では β=10^4 と非常に大きく設定してハードクラスタリングとすることで β の影響を受けないようにしているが，iIB 法および提案手法では結果は β の値に依存する．このため，予備実験により各手法に対する適切な β の範囲を求め，iIB 法では $\beta \in [1,100]$, 提案法では $\beta \in [10^{-2},1]$ として実験した．

スペクトルクラスタリングでは埋め込む次元数 l も影響を及ぼす．基本的に l=k(クラスタ数) としたが，Multi5 では 5 では低次元すぎると考え l=10 とした．

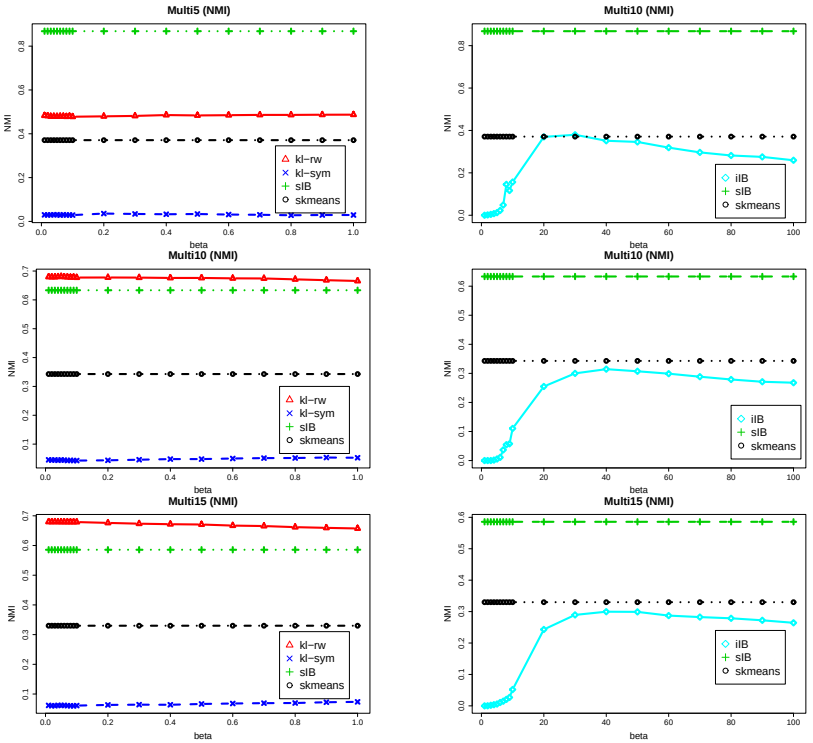

図 2 20NG に対する結果 (NMI)
Fig. 2 Result on 20NG (w.r.t. NMI)

4.3 結　　果

3 つの母集団に対してそれぞれ非復元抽出で 10 個ずつ作成した計 30 のデータセットに対し，初期値依存性を考慮して各データセットごとに 10 回試行を行った．各母集団に対する平均結果を図 2 に示す．図 2 ではグラフモデルに $\boldsymbol{L_{rw}}$ を用いたものを kl-rw, $\boldsymbol{L_{sym}}$ を用いたものを kl-sym と表記した．上記のように β が主要なパラメータであるため，各図ごとに横軸に β，縦軸に評価値とした．ただし，sIB 法では文献[8)] に従って $\beta = 10^4$ における各データセットに対する 10 回試行での最良値から平均を計算した．

各データのクラスタへの割り当ての正しさに対応する NMI に関しては (図 2)，Multi10, Multi15 に対して提案したグラフモデルに$\boldsymbol{L_{rw}}$を用いた手法 (kl-rw) が他手法より大きな値

Vol.2009-MPS-76 No.30
Vol.2009-BIO-19 No.30
2009/12/18

であった．他方，Multi5 に対しては，iIB 法や skmeans 法より大きな値であったが，sIB 法よりは小さかった．

4.4 考　　察

定理 1 での定常分布を求めるという意味では本稿での提案は iIB 法に対応すると考えられるが，4.3 節での結果より iIB 法との比較を通じて提案手法の妥当性と有効性を確認した．

本稿のアプローチはデータグラフ上の重みから導出される条件付確率に基づいて問題 1 をグラフ上の問題に帰着したものである．このため，条件付確率から導出されるグラフ上の酔歩に基づく$\boldsymbol{L_{rw}}$のほうが提案するグラフモデルに合致すると考えられる．図 2 の結果からも kl-rw が kl-sym を上回り$\boldsymbol{L_{rw}}$が提案モデルに合致することが確認できるため，グラフモデルは相互情報量に基づくクラスタリング問題のモデルとして妥当であると考えられる．

高次元スパースデータという困難な問題である文書クラスタリングに対して，提案法 (kl-rw) は NMI の評価で Multi10, Multi15 において sIB 法をも上回る性能を示し，クラスタ数が多い場合にクラスタへの割り当ての正しさという観点からの有効性を確認した．しかし，Multi5 に対しては残念ながら sIB 法には及ばなかった．この理由として，問題 1 は KL 情報量に基づく相互情報量により定式化されているが，文書クラスタリングへの適用に際してはテキスト処理でしばしば遭遇するゼロ頻度問題のために数値的に不安定となることが考えられる．実データへの適用に際して上記の問題に対処することは今後の課題である．

5. おわりに

本稿では，相互情報量に基づくクラスタリング問題に対するグラフモデルを提案し，データが一様分布する場合にはハードクラスタリング問題が提案するグラフモデルにおける組合せ最適化問題に近似できることを示した．相互情報量から導出される定常分布に着想を得たデータ間の類似度関数を定義してデータ集合全体を辺重み付きグラフとして表現し，もとの問題とグラフ上の組合せ問題との対応を示した．

提案するグラフモデルを用いてデータ集合を表現することにより，グラフ構造に基づく様々なアルゴリズムを用いて相互情報量に基づくクラスタリングを行うことが可能になると考えられる．一例として，文書クラスタリングでのベンチマークデータである 20 Newsgroup のデータをグラフモデルで表現し，スペクトルクラスタリング法を適用して評価し，他手法との比較を通じて提案手法の妥当性と有効性を確認した．

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) Cover, T. and Thomas, J.: *Elements of Information Theory*, Wiley (2006).
2) Dhillon, J. and Modha, D.: Concept decompositions for lage sparse text data using clustering, *Machine Learning*, Vol.42, pp.143–175 (2001).
3) Diestel, R.: *Graph Theory*, Springer (2006).
4) Ghosh, J.: *Scalable clustering*, pp.341–364, Lawrence Erlbaum Assoc. (2003).
5) Pereira, F., Tishby, N. and Lee, L.: Distributional clustering of English words, *Proc. of the 30th Annual Meeting of the Association for Computational Linguistics*, pp.183–190 (1993).
6) Ristad, E.: A Natural Law of Succession, Technical Report CS-TR-495-95, Princeton University (1995).
7) Slonim, N. and Tishby, N.: Agglomerative Information Bottleneck, *Advances in Neural Information Processing Systems (NIPS) 12*, pp.617–623 (1999).
8) Slonim, N.: The Information Bottleneck:Theory and Applications, PhD Thesis, Hebrew University (2002).
9) Slonim, N., Friedman, N. and Tishby, N.: Unsupervised Document Classification using Sequential Information Maximization, *SIGIR-02* (2002).
10) Tishby, N., Pereira, F. and Bialek, W.: The Information Bottleneck Method, *Proc. of the 37th Allerton Conference on Communication and Computation* (1999).
11) von Luxburg, U.: A Tutorial on Spectral Clustering, *Statistics and Computing*, Vol.17, No.4, pp.395–416 (2007).