

正規化最尤符号化に基づく グラフクラスタリングの研究

東京大学大学院情報理工学系研究科数理情報学専攻

平井聡 富岡亮太 山西健司

内容

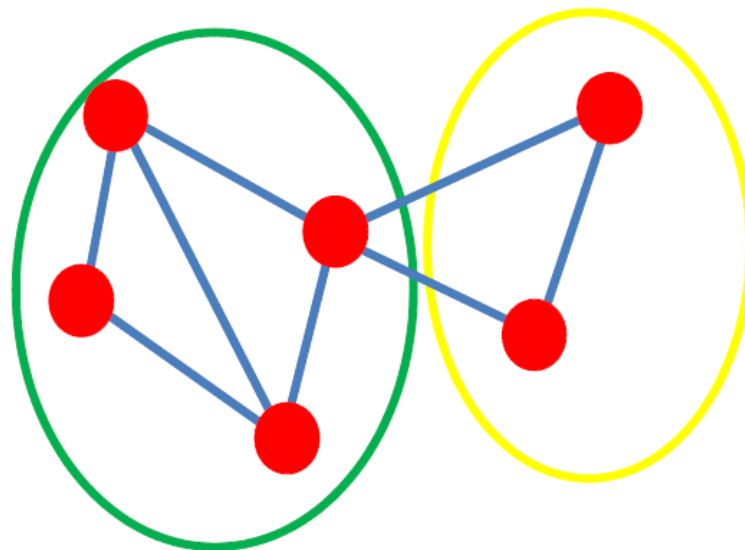
1. 問題提起
 グラフクラスタリングとは？
2. 目的
 正規化最尤符号化をグラフクラスタリングに適用
3. 評価基準
 正規化最尤符号長とは？
4. 提案手法
 グラフの構造にナイーブベイズモデルを導入
5. 実験考察
 人工データ (AICやBICとの比較)
 ベンチマークデータ
6. 結論

問題提起(1)

本研究におけるグラフクラスタリングとは
接続関係の与えられたグラフのノードを
疎密構造をもとにグループに分ける問題

例えば,

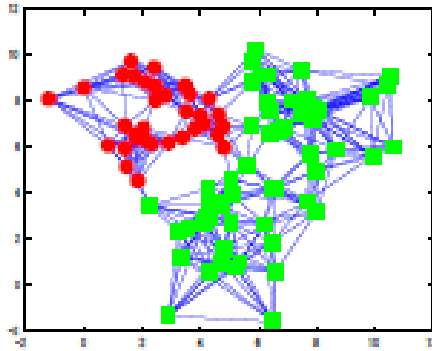
1. 購買に対する分割
2. 参照関係による論文の分割
3. 対人関係による分割



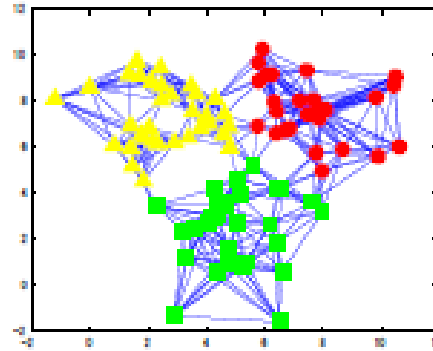
問題提起(2)

グラフクラスタリングの問題点

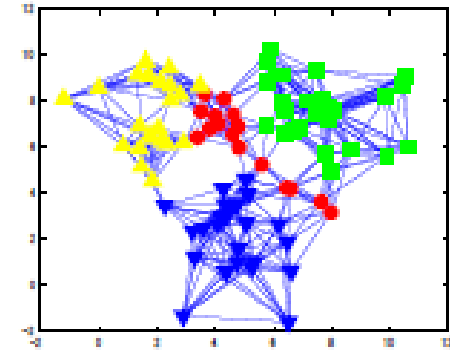
クラスター数の決定問題は一般に難しい



クラスター数：2



クラスター数：3



クラスター数：4

●情報量規準に基づく通常の性能評価規準:

- ・AIC (Akaike's Information Criterion) (H. Akaike, 1974)
- ・BIC (Bayesian Information Criterion) (G. Schwarz, 1978)

⇒これらと記述長最小化規準 (MDL規準) とを比較することを考える.

目的

本研究での性能評価

MDL原理の適用に正規化最尤 (NML) 符号化を使ったときにAICやBICと収束性などを比較

- MDL (Minimum Description Length) 原理に基づく性能評価:
データとモデルの総記述長が最小となるときにクラスター数が最適
- NML (Normalized Maximum Likelihood) 符号化:
もっとも短い記述長の計算方法
(Shtarkov, 1987, J. Rissanen, 1996)

ところが, これは一般に計算量が多く, 計算困難

⇒最近, 特別な場合 (ナイーブベイズモデルなど) には,
効率的な計算方法が提案されている
(P. Kontokanen & P. Myllymaki, 2009)

評価規準：NML符号長

NML符号長

MDL原理において、

Shtarkovのmin maxの意味でもっとも短い符号長

モデル: $\mathcal{M} = \{P(X|\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$

$$NML(\mathbf{x}^n | \mathcal{M}) = -\log \frac{P(\mathbf{x}^n | \hat{\theta}(\mathbf{x}^n, \mathcal{M}))}{\mathcal{C}(\mathcal{M}, n)}$$

最尤推定量

正規化項

$$= -\log P(\mathbf{x}^n | \hat{\theta}(\mathbf{x}^n, \mathcal{M})) + \log \mathcal{C}(\mathcal{M}, n)$$

ただし、

$$\mathcal{C}(\mathcal{M}, n) = \sum_{\mathbf{y}^n \in \mathcal{Y}^n} P(\mathbf{y}^n | \hat{\theta}(\mathbf{y}^n, \mathcal{M}))$$

⇒ 正規化項 $\mathcal{C}(\mathcal{M}, n)$ は計算困難

(例: 文字数が K の多項分布では、

$O(K^n)$ の計算量 n : データ数, K : 文字数)

⇒ $O(n^2 \cdot K)$ (n : データ数, K : クラスタ数)

ナイーブベイズモデル (P. Kontokaneni, P. Myllymaki, 2005)

提案手法(1)

NMLをグラフクラスタリングに適用 データの特徴ベクトルを接続行列で表現

手法のポイント

- ① 各データの特徴ベクトルをノード同士の接続行列で表現.
- ② NML符号長の計算にクラスターの分割を表す隠れ変数 Z を用いた.
(Z は与えられたものとする.)

これによって, AIC, BICにも適用可能.

$X =$

1	1	1	0	0	0	0
1	1	0	1	1	0	1
1	0	1	0	1	1	0
0	1	0	1	0	0	1
0	1	1	0	1	1	0
0	0	1	0	1	1	1
0	1	0	1	0	1	1

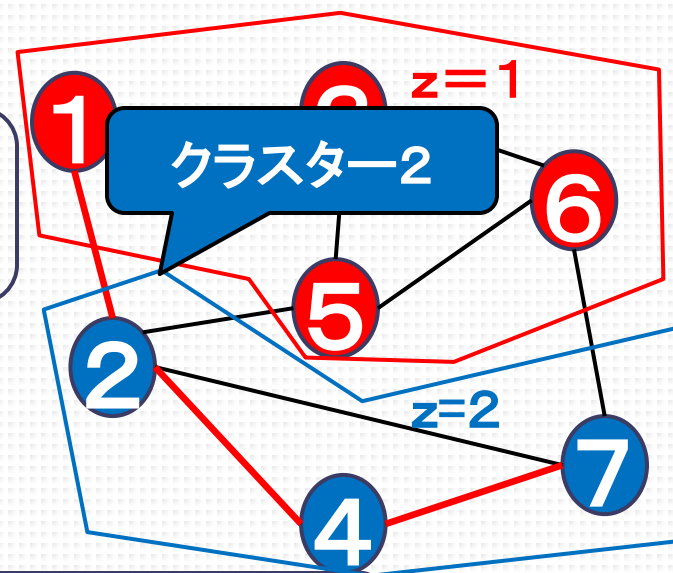
接続行列

←ノード1
←ノード2
←ノード3
←ノード4
←ノード5
←ノード6
←ノード7

ノード1の
ベクトル

ノード4の
ベクトル

ノード2はクラスター2に所属



提案手法(2)

グラフクラスタリングへの適用 グラフの構造にナイーブベイズモデルを導入

ナイーブベイズモデルとは？

- ① 各ノードの特徴ベクトルの要素はクラスター z にのみ依存し、
それぞれ独立：

$$P(z, x_1, \dots, x_n | \theta) = P(z | \theta) \cdot \prod_{i=1}^n P(x_i | z, \theta)$$

- ② 特徴ベクトルの要素 x は、接続関係であるので $\{0, 1\}$ をとる

$$P(x_j = 1 | z_k) = \sigma_{jk1}, \quad P(x_j = 0 | z_k) = \sigma_{jk0}$$

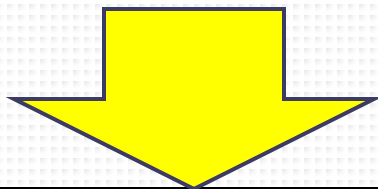
⇒このナイーブベイズモデルをグラフの構造に導入.

実験考察(1:人工データ)

グラフクラスタリングにおけるNML
NMLはAIC, BICより速い収束性を示す

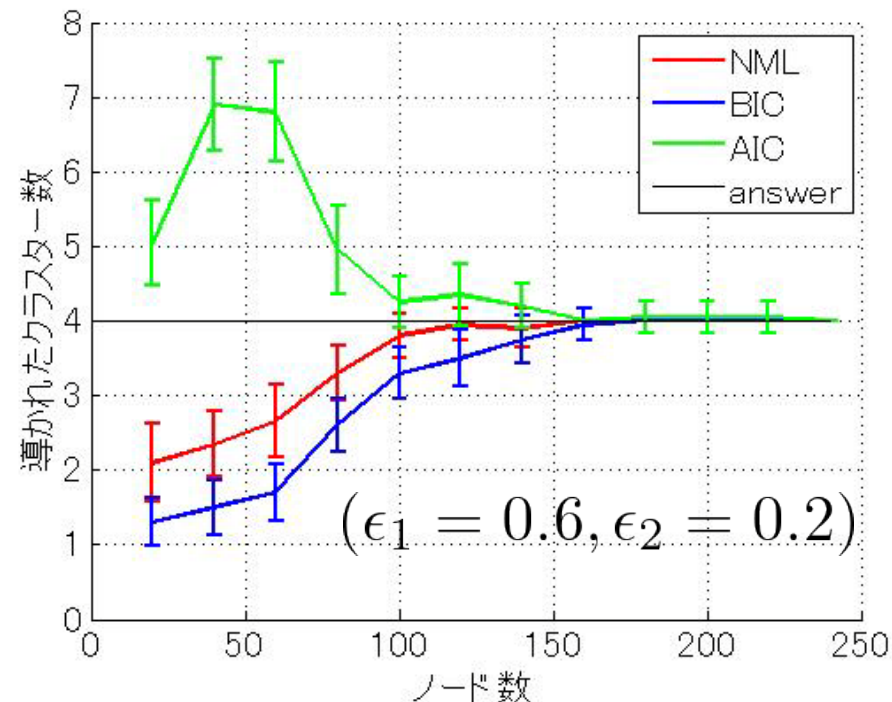
データ:

- ・真のクラスター数は3, 4, 5とする
- ・クラスター内とクラスター外のノードの接続確率をそれぞれ決める
- ・隠れ変数ZはK. Yu et. al. のクラスタリングアルゴリズムで求める

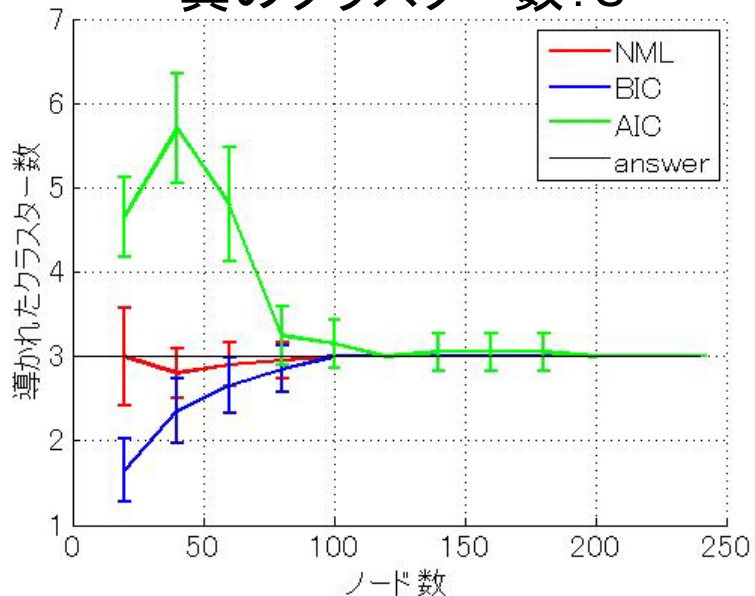


- グラフクラスタリングにおいて, NML, AIC, BICは収束性を示した
- NMLはこれらの中で最も収束が速い

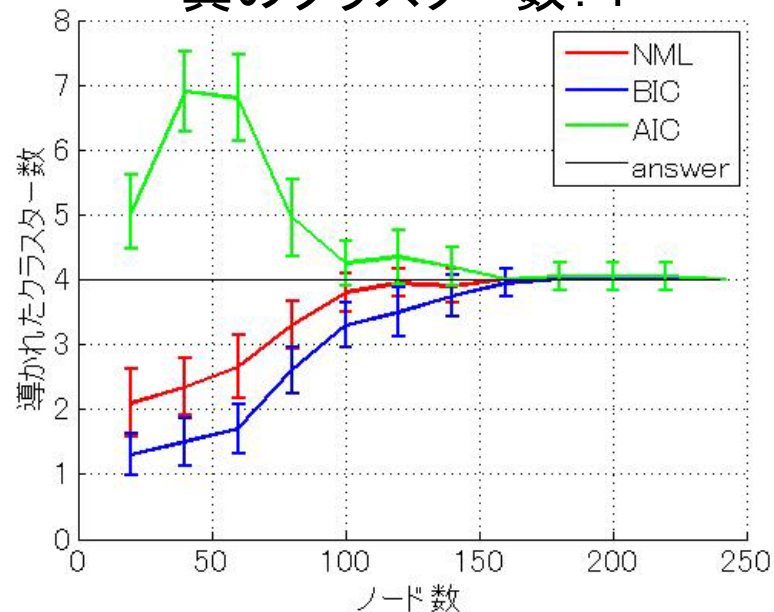
真のクラスター数: 4



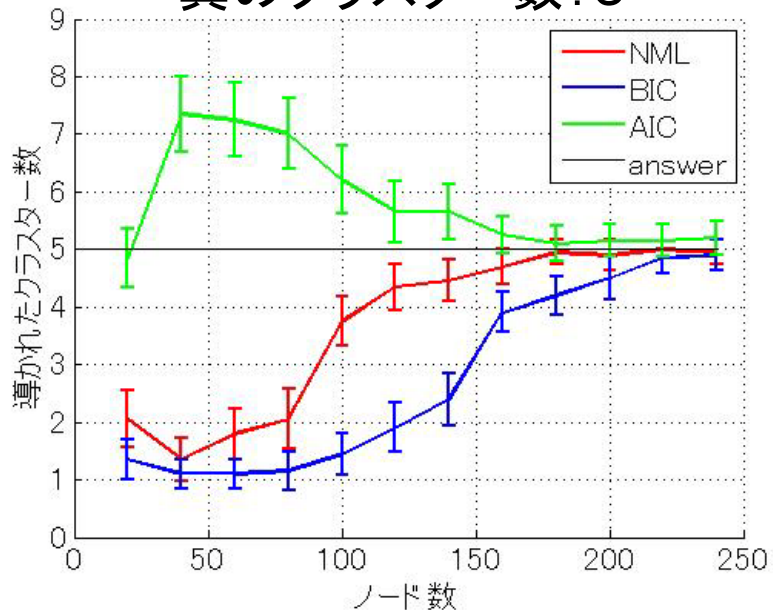
真のクラスター数: 3



真のクラスター数: 4



真のクラスター数: 5



真のクラスター数が3, 4, 5のときの
ノード数と導かれたクラスター数との関係
($\epsilon_1 = 0.6, \epsilon_2 = 0.2$)

いずれの場合もNMLが最も早く収束している
ことが分かる

実験考察(2:人工データ)

グラフ構造による収束性

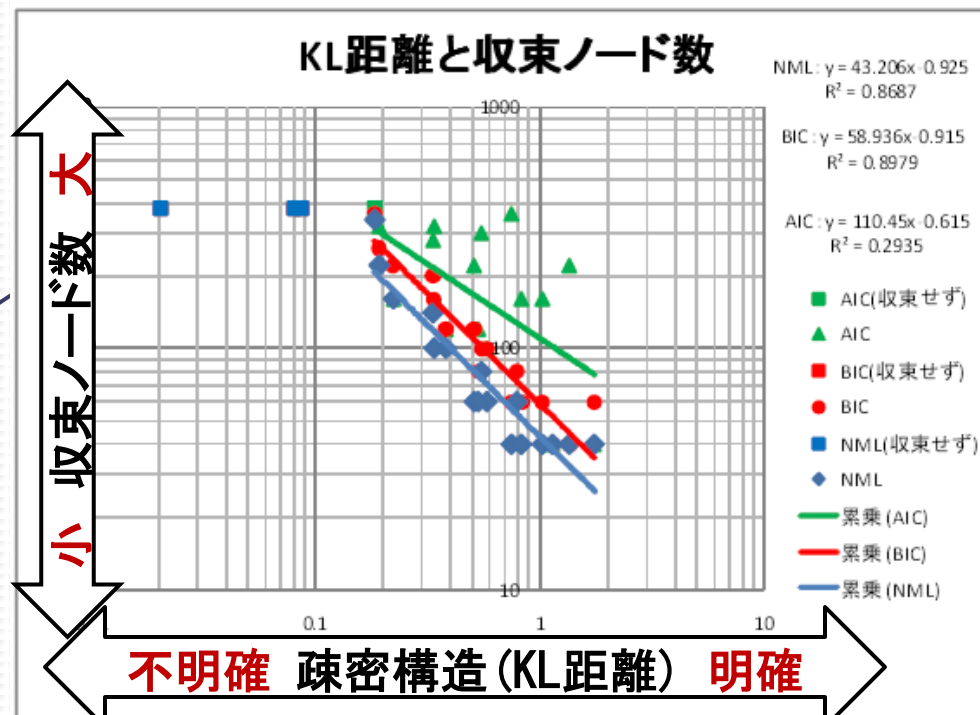
クラスター内外の疎密構造が明確なほど収束が速い

クラスター内のエッジの接続確率分布 ($P(\text{接続}) = \epsilon_1$) と
 クラスター外のエッジの接続確率分布 ($P(\text{接続}) = \epsilon_2$) の間の
 KL (Kullback-Leibler) 距離:

$$D(\epsilon_1 || \epsilon_2) = \epsilon_1 \cdot \log \frac{\epsilon_1}{\epsilon_2} + (1 - \epsilon_1) \cdot \log \frac{1 - \epsilon_1}{1 - \epsilon_2}$$

KL距離と収束ノード数:
 NML, BICにおいて, 実験的におよそ反比例であることが分かった

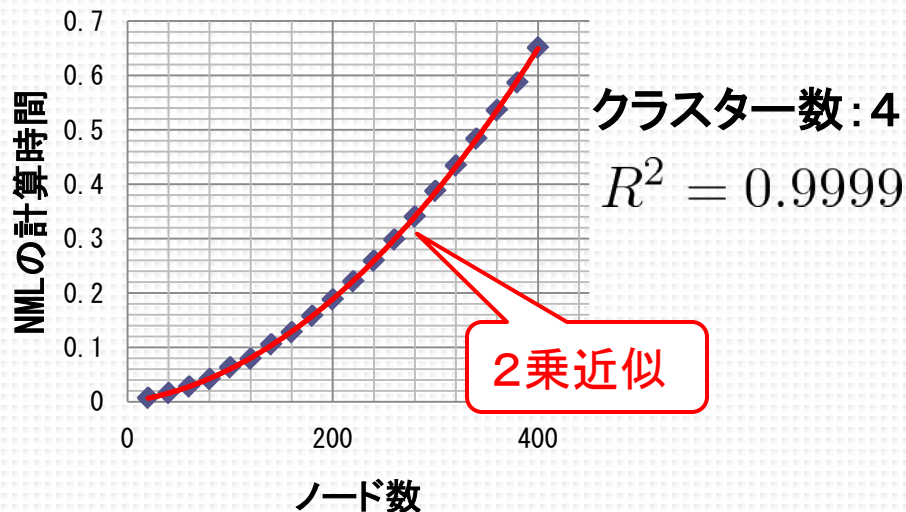
ただし, 収束ノード数とは
 収束までに必要な最小ノード数である.



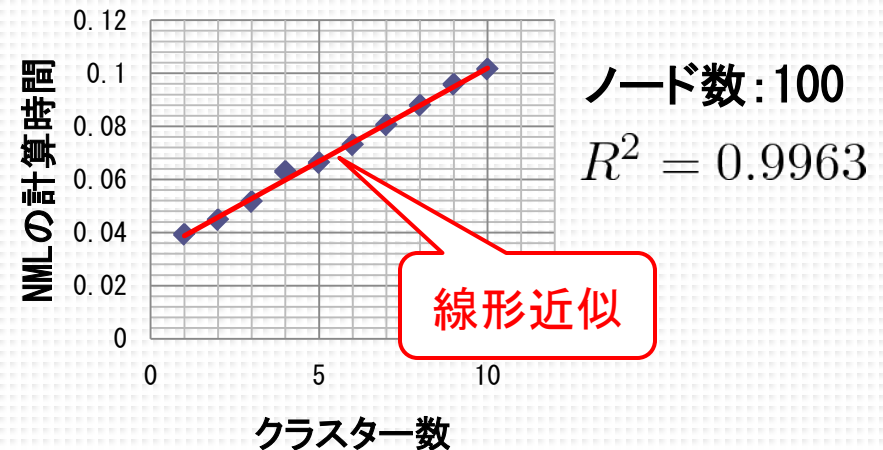
実験考察(3:人工データ)

計算時間

ノード数(n)の依存性



クラスター数(K)の依存性



理論的には計算時間は $O(n^2 \cdot K)$

実験的にも計算時間が $O(n^2 \cdot K)$ となることが確かめられた

実験考察(4:ベンチマークデータ)

ベンチマークデータ(レ・ミゼラブル)

物語の部の数(5)とクラスター数(4.66)がほぼ一致

評価基準	NML	BIC	AIC
平均	4.6600	2.7400	7.0400
標準偏差	0.3746	0.3050	0.3901

NMLが物語の部の数に最も近い

(クラスタリングの初期値を変えて50回実験

導かれたクラスター数の平均と標準偏差を表す)

データ:レ・ミゼラブル(データ数:77)

同じ章に出てきた人物同士にエッジが存在(章の数:365章)

(http://www.casos.cs.cmu.edu/computational_tools/datasets/external/lesmis/)

結論

NMLを用いたグラフクラスタリング

- グラフにナイーブベイズモデルを導入してNMLの効率的な計算方法を適用可能にした
- NMLはAIC, BICの規準よりも収束速度は速い
- クラスタ内外の疎密構造が明確なほど収束が速い

今後の研究

- 他の評価方法との比較.
- NMLの連続値データへの応用.