クラスタ構造を仮定した場合の双クラスタリングアルゴリズムの解析

Analysis of Biclustering Algorithms Assuming Several Types of Cluster Structure

山浦智佳子 [1*] Chikako Yamaura[1]
小林靖明 [1] Yasuaki Kobayashi[1]
山本章博 [1] Akihiro Yamamoto[1]
久保山哲二 [2] Tetsuji Kuboyama[2]

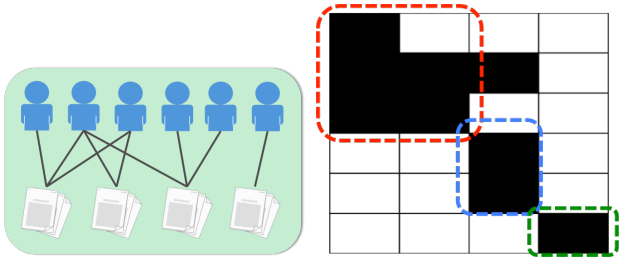
[1] 京都大学情報学研究科
[1] Graduate School of Informatics, Kyoto University
[2] 学習院大学計算機センター
[2] Computer Center, Gakushuin University

Abstract: Biclustering is a technique to extract dense submatrices in relational data represented as a matrix. It is recently used as graph clustering, collaborative filtering and micro-array data analysis. In biclustering we must consider several types of cluster structure behind an input data, where the structure in this research means the relation among biclusters. There are many biclustering algorithms proposed in the literature. Each algorithm extracts a set of biclusters with a specific structure. When the structure extracted by an algorithm does not match the desired structure, the algorithm may find clusters which is far from the desired communities. The structure extracted by algorithms do not necessarily match the desired structure. In this research, we formulated five types of bicluster structure and performed experiments to analyze behaviors of four biclustering algorithm.

1 はじめに

文章と単語の関係，著者と出版物の関係などを表す関係データには通常，互いに強く関連したコミュニティ構造が含まれている．このようなコミュニティは双クラスタリングにより見つけ出すことができる．双クラスタリングはデータにおける二つの属性を同時にクラスタリングする手法であり，テキストマイニング，協調フィルタリング，遺伝子データ解析などに広く用いられている．

本研究では特に二値行列で表すことのできる関係データの双クラスタリングについて扱う．図 1 は著者と出版物の関係における例である．図 1(a) で表される関係データは図 1(b) のような二値行列で表すことができる．行が著者，列が出版物，行列の内容はその著者がその出版物を書いたという関係を表している．また図 1(b) の赤，青，緑で囲われた部分は著者と出版物の関係が密に繋がっており，共著コミュニティを表している．このように関係データを表した二値行列における密な部分行列を双クラスタと呼び，双クラスタの集合を探し出すことを双クラスタリングと呼ぶ．



(a) 関係データ　(b) 二値行列による表現

図 1: 著者と出版物の関係データ

双クラスタの集合は，双クラスタ間の位置関係の制約によっていくつかの種類に分けることができる．図 1(b) では 3 つの双クラスタは行や列をお互いに共有しておらず，行，列ともに排他的な構造を取っている．しかしそれ以外にも，列のみを共有することを許す双クラスタ集合，重なりを許す双クラスタ集合などの場合が考えられる．このような双クラスタ集合における制約関係を本稿では双クラスタ構造と呼ぶ．

双クラスタ集合は図 2 に示す構造のうちいずれか一

*連絡先： 京都大学大学院情報学研究科
〒 606-8501 京都府京都市左京区吉田本町 36-1
E-mail: c.yamaura@iip.ist.i.kyoto-u.ac.jp

(a) (b) (c) (d) (e)

図 2: 双クラスタ構造．(a) 両側排他的構造，(b) 片側排他的構造，(c) 非排他的構造，(d) 重複構造，(e) 市松模様構造

つをとると考えることができる．コミュニティを含んでいるような関係データは本質的にこのいずれかの双クラスタ構造を持っているはずである．一方，それぞれの双クラスタリングアルゴリズムは，与えられた関係データの本来のクラスタ構造に関わらず，いずれかのクラスタ構造を抽出する．

本研究は，隠されたクラスタ構造を持ったデータを与えられたとき，いくつかの双クラスタリングアルゴリズムがその構造をどの程度上手く再現できるかを明らかにすることを目的とする．本稿では，双クラスタリングアルゴリズムとして両側排他的構造を持つ双クラスタ集合を抽出する二部モジュラリティ最適化，市松模様構造を抽出する符号化コスト最適化，重複構造を抽出する Bimax の拡張アルゴリズム，そして双クラスタリングの前処理として使うことのできる二部グラフ研磨アルゴリズムを扱い，これら四つのアルゴリズムがどのようなクラスタ構造のデータを与えられたときにどの程度再現ができるかを実験により確かめる．

本稿は以下の通り構成する．二章では準備として関係データやその表現，双クラスタ，双クラスタ構造について詳細な説明と定義を与える．また本稿と関連の深い先行研究についても触れる．三章では，本稿で扱う三つの双クラスタリングアルゴリズムと一つの前処理アルゴリズムについて説明する．四章で実験について，データの作成や評価方法，結果，考察を述べ，五章で本稿のまとめを行う．

2 準備

2.1 関係データと双クラスタ

本稿で扱うデータは二値関係データである．二値関係データは二つのオブジェクト集合 $X = \{x_1, ..., x_{n_R}\}$, $Y = \{y_1, ..., y_{n_C}\}$ とそれらの関係 $E \subseteq X \times Y$ を用いて (X, Y, E) で表される．ここでは X, Y, E はいずれも空集合ではないことと，X,Y の全ての要素が E に一度以上現れることを仮定する．すなわち $X, Y, E \neq 0$ かつ，任意の $x \in X$ について $(x, y) \in E$ となる $y \in Y$ が存在し，任意の $y \in Y$ についても $(x, y) \in E$ となる $x \in X$ が存在する．

実際に二値関係データを扱うときは，二値行列または二部グラフに変換して考える．二値行列は 0 または 1 のみを値としてとる行列である．本稿では二値行列を $A = (a_{ij})_{n_R \times n_C}$ で表し，$(x_i, y_j) \in E$ のときに限り $a_{ij} = 1$, そうでなければ $a_{ij} = 0$ をとるものとする．

双クラスタ集合は $\mathcal{B} = \{B_1, ..., B_k\}$ で表される．ここで i 番目の双クラスタは行クラスタ $R_i \subseteq \{1, ..., n_R\}$ と列クラスタ $C_i \subseteq \{1, ..., n_C\}$ の対であり，$B_i = (R_i, C_i)$ と表す．二値行列表現では，双クラスタは元の行列の部分行列となる．双クラスタは通常，十分に密である（1 の比率が高い）ような部分が要求される．ただし詳細な定義は扱いたいコミュニティの性質やアルゴリズムに依存する．

二値関係データは（重みなし）二部グラフでも表現することができる．二部グラフは各頂点が共通部分を持たない二つの集合に分割され，同じ集合内の頂点間には辺がないようなグラフである．本稿では二部グラフは上述の X, Y, E を用いて $G = (X, Y, E)$ で表される．二部グラフ表現においては双クラスタは元のグラフの部分グラフとなる．

二値行列と二部グラフは同じ対象を指しており，互いに変換可能である．本稿では二値行列表現と二部グラフ表現の両方を区別なく用いる．

2.2 双クラスタ構造

双クラスタの集合は，それらに含まれるクラスタ間の制約関係によっていくつかの種類に分けることができる．本稿では Madeira ら [1] に従ってこのクラスタ間の制約関係の種類のことを双クラスタ構造と呼ぶ．本稿では以下の 5 つの双クラスタ構造を与える．

両側排他的構造 任意の二つの双クラスタにおいて，その行クラスタ間にも列クラスタ間にも共通部分がないような構造である．すなわち $i, j \in \{1, ..., k\}, i \neq j$ について $R_i \cap R_j = \emptyset$ かつ $C_i \cap C_j = \emptyset$ である．

片側排他的構造 列排他的構造と行排他的構造がある．列排他的構造は任意の二つの双クラスタにおいて，列クラスタ間は共通部分を持たないがが行クラスタ間は共通部分を持つことが許されるような構造である．すなわち $C_i \cap C_j = \emptyset$ である．行排他的構造についても同様に定義される．両者には本質的な違いはないため，本稿では列排他的構造のみを扱う．また両側排他的構造はこの構造の特殊な場合と考えることができる．

非排他的構造 行クラスタ間，列クラスタ間の両方において共通部分を持つことが許されるが，双クラスタ間の重なりは許されない．$B_i \cap B_j = \emptyset$ で定義

される．片側排他的構造はこの構造の特殊な場合と考えることができる．

重複構造 双クラスタ間の重なりが許される．ただし本稿ではある双クラスタが完全に他の双クラスタに含まれることは許さないものとする．$B_i \backslash B_j \neq \emptyset$ で定義される．非排他的構造はこの構造の特殊な場合と考えることができる．

市松模様構造 排他的な行のクラスタ集合 $\mathcal{U}$ と排他的な列のクラスタ集合 $\mathcal{V} = \{V_1, ..., V_{k_C}\}$ を独立して扱うような構造である．双クラスタ集合 $\mathcal{B}$ はそれらの直積として与えられる．すなわち行において $\mathcal{U} = \{U_1, ..., U_{k_R}\}, U_1 + ... + U_{k_R} = \{1, ..., n_R\}$ であり，また列において $\mathcal{V} = \{V_1, ..., V_{k_C}\}, V_1 + ... + V_{k_C} = \{1, ..., n_C\}$ であり，双クラスタ集合は $\mathcal{B} = \mathcal{U} \times \mathcal{V}$ となる．この構造においては，二値行列の全ての要素が一つの双クラスタに属するため，全ての双クラスタが密であるわけではないことに注意する．通常この構造を扱うアルゴリズムでは，それぞれの双クラスタが十分に密か十分に疎かのいずれかとなるように目的関数を設定する．

本稿では関係データは本来上記のいずれかのクラスタ構造を本質的に持っていると仮定する．一方双クラスタリングアルゴリズムもまた上記のいずれかの構造のクラスタ集合を抽出する．

2.3 関連研究

本稿と関連が深い先行研究について述べる．クラスタ集合の評価尺度であるモジュラリティ [7] の最適化によるクラスタリング手法は二部グラフに限らず一般的なグラフネットワークからコミュニティを抽出する手法として広く利用されている．二部グラフに特化したモジュラリティとしては本稿で扱う Barber [2] のもの以外にも市松模様構造を扱う Suzuki [8] のものなど複数が提案されている．

モジュラリティ以外の評価尺度を用いた手法もある．本稿で扱う Gao [3] らによる符号化コストは情報理論における最小記述長 (MDL) の原則に基づいている．

Madeira ら [1] は本稿で与えた 5 つを含む 9 つの双クラスタ構造を与え，既存の双クラスタリングアルゴリズムがどの構造を想定しているかを述べた．ただし彼らの研究対象は遺伝子発現データに対する双クラスタリングであり，本稿で扱う二値関係データのコミュニティとは異なる性質を持っている．またデータが本来持っているクラスタ構造という観点については言及していない．

3 アルゴリズム

3.1 二部モジュラリティ最適化

モジュラリティはグラフに対するクラスタ集合の質を評価する尺度であり，二部グラフに限らずネットワークのコミュニティ抽出に広く利用されている．扱うグラフやコミュニティの性質に合わせた様々な種類のモジュラリティが提案されているが，ここでは二部グラフとその頂点の分割に対する質を測るために考案された Barber のモジュラリティ [2] を扱う．

二部グラフ $G = (X, Y, E)$ と両側排他構造をとる双クラスタ集合 $\mathcal{B}$ に対し Barber のモジュラリティ Q は以下で与えられる．

$$Q = \sum_{(R_i, C_i) \in \mathcal{B}} \left(\frac{2|R_i \to C_i|}{|X \to Y|} - \frac{|R_i \to Y||C_i \to X|}{|X \to Y|^2} \right)$$

ここで頂点集合 S, T に対し $|S \to T|$ は S 内の頂点から T 内の頂点へ繋がる辺の総数を表す．ただし自己ループ，すなわち両端が S と T の共通部分にあるような辺は二回ずつ数えることに注意する．

モジュラリティは高い値であるほどそのクラスタ分割が良質であることを表し，常に $Q \leq 1$ を満たし，また全ての頂点が同じクラスタに含まれる場合に $Q = 0$ となる．

クラスタリングはモジュラリティが極大な値をとるクラスタ分割を探索することで行うことができる．本稿では Barber のモジュラリティ最適化手法として Louvain 法 [4] を用いた．

3.2 符号化コスト最適化

符号化コストはモジュラリティと同じく双クラスタ集合の質を評価する尺度である．

本稿では Gao ら [3] による定義を用いる．二値行列と市松模様構造をとる双クラスタ集合が与えられたとき，それらに関する情報を可逆符号化することを考える．このとき符号化すべき情報は，二値行列のサイズ，行クラスタと列クラスタの数，各行各列からクラスタへのマッピング，各双クラスタにおける 1 の数，各双クラスタ内の実際の行列の値である．

したがって双クラスタ集合を $\mathcal{B} = \mathcal{U} \times \mathcal{V}$ とし，二値行列を二部グラフ $G = (X, Y, E)$ で表すとき，符号化コスト L は以下で与えられる．

$$\begin{aligned} L \quad = \quad & \log^* |X| + \log^* |Y| + \log^* |U| + \log^* |V| \\ & + \sum_{U_i \in \mathcal{U}} |U_i| \log \left(\frac{|X|}{|U_i|} \right) + \sum_{V_j \in \mathcal{V}} |V_j| \log \left(\frac{|Y|}{|V_j|} \right) \end{aligned}$$

$$+\sum_{U_i\in\mathcal{U},\,V_j\in\mathcal{V}}\left(\log\left(|U_i||V_j|+1\right)\right.$$
$$\left.+|U_i||V_j|H\left(\frac{|U_i\to V_j|}{|U_i||V_j|}\right)\right).$$

ただし $\log^* n$ は $\log n$, $\log\log n$, ... を正の項について足し合わせたもの，H はエントロピー関数 $H(p) = -p\log p-(1-p)\log(1-p)$ である．

L は常に 0 より大きな値をとる．最小記述長 (MDL) の原則に基づき，この符号化コストが小さいほど良質なクラスタ分割であると考えられ，クラスタリングはこのコストが極小な値をとる分割を探索することで行う．本稿では Gao らのアルゴリズムを用いた．

3.3 拡張 Bimax

Bimax は二部グラフから極大な二部クリークを全て抽出するアルゴリズムである [5] ここで二部グラフ G における極大な二部クリーク $G'=(X',Y',E')$ は，以下で定義される．(1) G' は G の部分グラフである．(2) 任意の X' 内頂点と Y' 内頂点の間に辺がある．すなわち任意の $x\in X', y\in Y'$ において $(x,y)\in E'$．(3) 二部クリーク $G''=(X'',Y'',E''), X'\subseteq X'', Y'\subseteq Y''$ が存在するとき $G''=G'$ である．

Bimax は分割統治法を採用しており，与えられたグラフを部分グラフに分割しながら再帰的に極大二部クリークを探索する．しかし元々の Bimax アルゴリズムは全ての極大二部クリークを列挙するために計算時間が非常に長くなってしまうことがある．そのため我々はアルゴリズムを拡張し，現在注目している部分グラフに含まれる任意の辺が，既に抽出された二部クリークに含まれている場合は，この部分グラフの探索を中止するようにした．この拡張により双クラスタリングにおいて影響の少ないと思われる二部クリークの探索を省略することができ，計算時間が大幅に短縮される．

極大二部クリークは (上記の方法で列挙を省略した場合でも) お互いに重複が可能であるため，このアルゴリズムで得られる双クラスタは重複構造をとる．

3.4 二部グラフ研磨

双クラスタリングアルゴリズムの他に前処理アルゴリズムとして二部グラフ研磨 [6] を扱う．グラフ研磨は与えられたグラフの密な部分と疎な部分を強調し，グラフの特長を明確化するアルゴリズムである．小規模で密なクラスタを発見するために開発されたもので，クリークの列挙と相性が良い．

二部グラフ $G=(X,Y,E)$ とパラメータ σ_1, σ_2,T が与えられたとき以下の手順を行う．

(1) 新しく空の二部グラフ $G'=(X',Y',E')$ s.t. $X',Y',E'=\emptyset$ を作成．(2) $N(x)$ を x の隣接頂点, $sim(S,T)$ を頂点集合 S,T 間の Jaccard 距離とする．各 $x\in X$ について, x と似ている X 上の頂点集合 $S=\{x'\in X|sim(N(x),N(x'))\le\sigma_1\}$ を作成する．次に S に類似した Y 上の頂点集合 $T=\{y\in Y|sim(N(y),S)\le\sigma_2\}$ について x から各 $y\in T$ への辺を G' に追加する．(3) G と G' が同じであるかまたは反復回数が T に達した場合 G' を返し，そうでなければ $G=G'$ とし，手順 1 に戻る．

4 実験

それぞれのアルゴリズムが，関係データが持つ隠れた双クラスタ構造をどの程度復元できるかを見るため，人工データを作成し実験を行った．

4.1 データ作成

作成するデータは二値行列と正解の双クラスタ集合を含む．以下に示す 6 つのパラメータを用いてデータの作成を行う．(1) 行列の行のサイズ n_R (100 に固定)．(2) 行列の列のサイズ n_C (100 に固定)．(3) 双クラスタ構造の種類．(4) 作成する双クラスタ集合の基盤となるグループの数 g (10 に固定)．(5) 構造の強さの程度を表すパラメータ p．(6) ノイズ発生確率 ϵ．

データの作成は以下の手順で行う．(1) $n_R\times n_C$ サイズの空の行列を作成する．(2) クラスタ構造と p に従って全ての双クラスタの位置を決める．(3) 双クラスタの位置と ϵ に従って行列内の実際の値を決定する．

(2) の双クラスタの位置決めでは，最初に両側排他的構造をとるクラスタ集合を作成してから p に従い個別の構造の特長を強くしていく．まず片側排他的構造の場合を述べる．行と列をそれぞれ g 個の同じサイズの素集合に分割し，行の k 番目の素集合を U_k, 列の l 番目の素集合を V_l, i 番目の行を r_i, j 番目の列を c_j とする．次に対角線上にできた長方形を双クラスタとする．すなわち $s\in\{1,...,g\}$ について $B_s=(U_s,V_s)$ とし，$\mathcal{B}=\{B_1,...,B_g\}$ とする．ここで $\mathcal{B}$ は両側排他的な双クラスタ集合となっている．次に (r_i,V_l) で表される全ての領域のうち，まだいずれの双クラスタにも含まれていないものをランダムな順番で選択し，B_l に追加していく．双クラスタ集合の被覆率が p を超えない間この手順を続ける．ここで双クラスタ集合 $\mathcal{B}$ による被覆率を，$\mathcal{B}$ の総面積から初期状態の $\mathcal{B}$ の面積を除いた面積の行列サイズにおける割合で定義する．すなわち $\sum_{B_i\in\mathcal{B}}|R_i||C_i|-\sum_{k=1}^{g}|U_k||V_k|$ である．

以上により片側排他的構造を持つクラスタ集合 $\mathcal{B}$ が得られる．次に手順 (3) により，行列の各要素について

$\mathcal{B}$ に含まれれば確率 $1-\epsilon$, 含まれなければ確率 ϵ で 1 とする．以上により二値行列と正解クラスタが得られる．

非排他的構造，重複構造も作成方法も同様である．ただし (r_i, V_l) に加えて (U_k, c_j) で表される領域も選択し，前者は B_l に追加され後者は B_k に追加される．さらに非排他的構造では，領域の追加後に双クラスタがお互い重複していないかを確認し，もし重複があれば直前の追加をキャンセルする．

市松模様構造では，行と列を分割した後，(U_k, V_l) で表される全ての領域を $\mathcal{B}$ とする．さらにその中での対角線上に存在する双クラスタ (U_s, V_s) を「黒」クラスタ，それ以外を「白」クラスタとする．次に白クラスタをランダムな順番で選び，全ての黒クラスタによる被覆率が p を超えない間，黒クラスタに変換する．$\mathcal{B}$ が得られたら黒クラスタに含まれる要素を確率 $1-\epsilon$, 白クラスタに含まれる要素を確率 ϵ で 1 とする．以上により全ての構造において二値行列と正解クラスタが得られる．

4.2 評価方法

得られたクラスタ集合の評価に NMI(正規化相互情報量 normalized mutual information) を用いる．本稿で扱う全ての双クラスタ構造について NMI を適用するために予めいくつかの変換を行う．まず二値行列の要素のうち値が 1 であるものをそれぞれ一つのデータ点とみなし，クラスタを（双クラスタではない）通常のクラスタと考える．クラスタ間に重複がある場合，各要素に対し代表クラスタを決めることで強制的にクラスタ間の重複をなくす．代表クラスタは，最も多くのデータ点を含むものを選ぶ．複数の候補がある場合，予めデータ点に番号を振っておき，最も若い番号のデータ点を含むクラスタを選ぶ．さらにどのクラスタにも含まれないデータ点それぞれに，新たなクラスタを一つずつ割り振る．以上により各データ点が丁度一つのクラスタに含まれることとなる．

N 個のデータセットとその正解クラスタ集合 $S = \{S_1, ..., S_k\}$, さらにアルゴリズムにより得られたクラスタ集合 $T = \{T_1, ..., T_l\}$ が与えられたとき，S, T 間の NMI は以下で求められる．

$$NMI(S,T) = \frac{I(S,T)}{(H(S)+H(T))/2}$$

ここで $I(X,Y) = \sum_i \sum_j P(X_i \cap Y_j) \log \frac{P(X_i \cap Y_j)}{P(X_i)P(Y_j)}$ は相互情報量，$H(X) = -\sum_i P(X_i) \log P(X_i)$ はエントロピー関数，さらに $P(X_i) = |X_i|/N$ はデータがクラスタ X_i に入る確率，$P(X_i \cap Y_j) = |X_i \cap Y_j|/N$ はデータがクラスタ X_i と Y_j の両方に入る確率を表す．

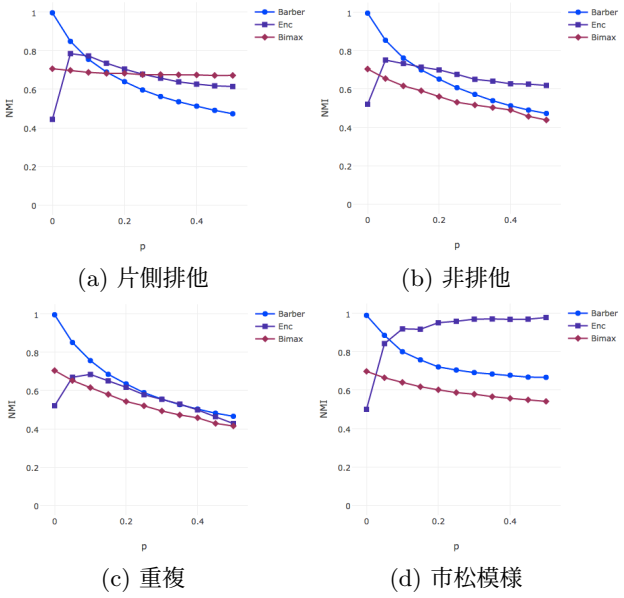
NMI は 0 以上 1 以下の値をとり，クラスタ S と T が完全に同じ場合 $NMI(S,T) = 1$ となる．

4.3 結果と考察

4 章で述べた 3 つの双クラスタリングアルゴリズムにつきそれぞれ研磨による前処理あり，なしを考慮した合計 6 種類のアルゴリズムについて，それぞれの双クラスタ構造を持つデータを与え，NMI を測定した．また測定の際は 10 回の実験の平均値をとった．

表 1: 結果概要

	両側	片側	非排他	重複	市松
Barber	◎	×	×	×	○
Enc	×	○	○	×	◎
Bimax	○	○	×	×	×
Barber+Polish	◎	×	○	×	○
Enc+Polish	◎	×	○	○	×
Bimax+Polish	◎	×	○	◎	×

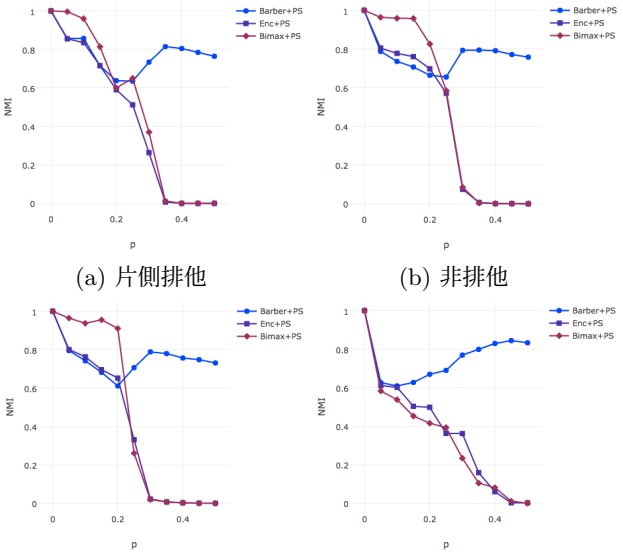


(a) 片側排他　(b) 非排他

(c) 重複　(d) 市松模様

図 3: 研磨なし

表 1 は結果の概要である．ここでは $p = 0.2$, $\epsilon = 0.05$ とし，NMI0.9 以上を◎， 0.7 以上を○， それ未満を×とした．さらに図 3， 4 は片側排他的構造，非排他構造，重複構造，市松模様構造について p を 0 から 0.4 まで変化させた場合の NMI の変化である．青，紫，赤の折れ線はそれぞれ Barber のモジュラリティの最適化 (Barber), 符号化コスト最適化 (Enc), 拡張 Bimax (Bimax) を表している．前節で述べた通り，片側排他，非排他，重複においては $p = 0$ に近いほど両側排他的構造に近いことを意味する．ただし市松模様構造には疎なクラスタも存在するため少し挙動が異なる．また Barber, Enc, Bimax はそれぞれ両側排他的構造，市松



(a) 片側排他　(b) 非排他

(c) 重複　(d) 市松模様

図 4: 研磨あり

模様構造，重複構造を抽出するアルゴリズムである．

まずアルゴリズムが想定するものと合致するクラスタ構造が与えられた場合，研磨 (Polish) なし Barber と研磨なし Enc は非常に高い精度を得ている事が表から見て取れる．一方 Bimax はクリークを抽出するためにノイズの影響を大きく受けてしまうが，研磨あり Bimax ではノイズの影響が小さくなり Bimax が想定する重複構造において高い精度を得ている．

また 3 種のアルゴリズムの中では，Enc は構造の変化に対して比較的堅牢であることもわかる．

次に入力が両側排他的構造に近いとき，すなわち片側排他，非排他，重複構造において p が 0 から 0.2 程度であるとき，研磨ありのアルゴリズムはいずれも高い精度を取っている．これは二部グラフ研磨が両側排他的構造を最も得意としているとも考えられる．しかし，p が 0.2 を超えたあたりから急激に精度が下がっていることから，ある時点に相転移が存在しているようである．

5　まとめ

本研究では 5 種類の双クラスタ構造に着目し，3 つの双クラスタリングアルゴリズムと 1 つの前処理アルゴリズムについてクラスタ構造の再現性を調べた．結果，各アルゴリズムは入力データが自らの想定と合致する構造を持っている場合は高い精度で再現が可能であること，また符号化コスト最適化アルゴリズムは入力の構造の変化に対して堅牢であること，さらに二部グラフ研磨アルゴリズムは両側排他的構造に近い入力を与えられた場合，精度を改善することが多いが，構造を変化させるとある時点で相転移が現れ，一気に精度が下がってしまうことが分かった．

また課題として，本稿では関係データに隠れた双クラスタ構造があることを仮定したが，未知の関係データが与えられたときにどの構造が潜んでいるかを判定する方法が必要であることが挙げられる．

謝辞

本研究は一部，JST，CREST および，JSPS 科研費 26280085，26280090 の支援を受けている．

参考文献

[1] Madeira, S. C. and Oliveira, A. L.: Biclustering algorithms for biological data analysis: a survey, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Vol. 1, No. 1, pp. 2445 (2004).

[2] Barber, M. J.: Modularity and community detection in bipartite networks, *Physical Review E*, Vol. 76, No. 6, p. 066102 (2007).

[3] Gao, T. and Akoglu, L.: Fast information-theoretic agglomerative coclustering, *Australasian Database Conference*, Springer, pp. 147159 (2014).

[4] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E.: Fast unfolding of communities in large networks, *Journal of statistical mechanics: theory and experiment*, Vol. 2008, No. 10, p. P10008 (2008).

[5] Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Büuhlmann, P., Gruissem, W., Hennig, L., Thiele, L. and Zitzler, E.: A systematic comparison and evaluation of biclustering methods for gene expression data, *Bioinformatics*, Vol. 22, No. 9, pp. 11221129 (2006).

[6] 中原 孝信, 大内 章子, 宇野 毅明, 羽室 行信, 「データ研磨の 2 部グラフへの適用と Twitter からの意見抽出」,2016 年度人工知能学会（第 30 回）, 北九州国際会議場, 2016.6.6〜6.9, 発表 6.2.

[7] Newman, M. E. and Girvan, M.: Finding and evaluating community structure in networks, *Physical review E*, Vol. 69, No. 2, p. 026113 (2004).

[8] Suzuki, K. and Wakita, K.: Extracting multi-facet community structure from bipartite networks, *Computational Science and Engineering, 2009. CSE '09. International Conference on*, Vol. 4, IEEE, pp. 312319 (2009).