b-彩色に基づくクラスタリングの色付け替え手法

吉田　哲也† HaythamELGHAZEL†† VéroniqueDESLANDRES††
Mohand-SaidHACID††† AlainDUSSAUCHOY††

† 北海道大学大学院情報科学研究科　〒 060–0814 札幌市北区北 14 条西 9 丁目
†† PRISMa Laboratory, Claude Bernard University of Lyon I　43 Bd du 11 novembre 1918, 69622 Villeurbanne cedex, France
††† LIRIS Laboratory, Claude Bernard University of Lyon I　43 Bd du 11 novembre 1918, 69622 Villeurbanne cedex, France
E-mail: †yoshida@meme.hokudai.ac.jp, ††{elghazel,deslandres,dussauchoy}@bat710.univ-lyon1.fr, †††mshacid@liris.cnrs.fr

あらまし　本稿では，b-彩色に基づくクラスタリング結果（データの分割）に対し，b-彩色という性質を満たしながら分割の質を向上する手法を提案する．b-彩色に基づくクラスタリング手法は，クラスタ数が指定されない場合でも，精緻な分割を与える手法として提案され，外れデータに対する頑健性など，クラスタリング手法としての望ましい性質を持つ．しかし，グラフの b-彩色に基づきクラスタリングを行う際，クラスタの質は明示的に考慮されていなかった．本稿で提案する手法ではこの点を補完し，b-彩色という性質を保証しながらクラスタの質の向上を実現する．提案する手法を UCI リポジトリのベンチマークデータに対して評価し，その有効を確認した．
キーワード　クラスタリング，b-彩色，色付け替え

A Re-Coloring Method for b-coloring based Clustering

Tetsuya YOSHIDA†, Haytham ELGHAZEL††, Véronique DESLANDRES††, Mohand-Said HACID†††, and Alain DUSSAUCHOY††

† Grad. School of Info. Science and Technology, Hokkaido University　N-14 W-9, Sapporo, 060–0814 Japan
†† PRISMa Laboratory, Claude Bernard University of Lyon I　43 Bd du 11 novembre 1918, 69622 Villeurbanne cedex, France
††† LIRIS Laboratory, Claude Bernard University of Lyon I　43 Bd du 11 novembre 1918, 69622 Villeurbanne cedex, France
E-mail: †yoshida@meme.hokudai.ac.jp, ††{elghazel,deslandres,dussauchoy}@bat710.univ-lyon1.fr, †††mshacid@liris.cnrs.fr

Abstract　This paper proposes a re-coloring method for b-coloring based clustering to improve the specified b-coloring partition while satisfying b-coloring property. The b-coloring based clustering method in [3] enables to build a fine partition of the data set in classes even when the number of classes is not pre-defined. However, it does not consider the *high quality* of the clusters in the construction of a b-coloring graph. The proposed method in this paper can complement its weakness by re-coloring the objects to improve the quality of the constructed partition under the property and the dominance constraints. The proposed method is evaluated against benchmark datasets and its effectiveness is confirmed.
Key words　clustering, b-coloring, re-coloring

1. はじめに

クラスタリングとは，与えられたデータの集合を，同じグループに属するデータ同士は類似し，異なるグループに属するデータ同士は類似しないようないくつかのグループ（クラスタ）に分割するプロセスである [5]．クラスタリングは，ウェブ解析，

情報検索などのにおいて重要な役割を果たす．様々なクラスタリング手法の概略については [9] を参照されたい．

与えられたデータに対し，全てのデータ対間の距離（あるいは，非類似度）が与えられた場合，各データ対 (v_i, v_j) に対して対応する非類似度 $D(v_i, v_j)$ を保持するような非類似度行列 D を用いて表現することができる．対象データにおける非類似度行列 D が与えられれば，個々のデータを頂点 v_i に対応させ，データ間の距離をラベル $d(v_i, v_j)$ とみなすことで，非類似度行列 D をグラフとして捉えることができる．

データ集合をグラフの観点から捉えてクラスタリングを行う手法として，グラフ理論に基づく手法がある [9]．このアプローチでは，基本的には類似度に基づいてグラフの中から組合せ論的な構造を探索する．グラフの彩色に基づくクラスタリング手法もいくつか提案されている．文献 [4] ではグラフの最小全域木に対する 2 彩色に基づいた階層的クラスタリング手法が提案されている．また，文献 [6] では，最小の直径を持つ k 個のクラスタへの分割問題を閾値グラフにおける最小彩色問題に還元した手法が提案されている．

近年，グラフ理論における b-彩色 [2], [8] に基づいたクラスタリング手法が提案された [3]．グラフの b-彩色とは，グラフの頂点を以下の制約を満たす最大の彩色数で頂点を彩色することである：

(i)　（辺で）隣接する任意の 2 頂点は異なる色を持つ (適正彩色)

(ii)　同じ色で彩色された頂点の集合が 1 つのクラスタに対応するが，各クラスタには，他の全て色の頂点に隣接する頂点（支配頂点）が少なくとも 1 つ存在する．

文献 [3] では，b-彩色を満たすグラフを構築する手法が提案された．しかし，この手法では b-彩色性を満たすグラフを構築することはできるが，グラフを構築する際には基本的には b-彩色性を満たすことのみが考慮され，クラスタの質は明示的に考慮されていなかった．

本稿では，b-彩色性を満たしながら分割（クラスタリング結果）の質を向上させる色付け替え手法を提案する．本稿で提案する手法は，b-彩色性を満たす分割の質を向上することができるため，文献 [3] での手法を補完する役割を果たす．提案手法では，分割における支配頂点の色に影響を与えない頂点を選択し，分割の質を単調に増加させるながら頂点の色を付け替える．前者により色付け替えを行っても適正彩色と支配頂点という b-彩色性を保つことを保証し，後者により色付け替え後のクラスタの質を向上させることを保証する．提案手法を UCI リポジトリ [7] のベンチマークデータに対して評価し，その有効性を確認した．

2. 節では，文献 [3] で提案したクラスタリング手法の概要を説明し，その挙動を例を用いて説明する．クラスタリングの妥当性指標に基づく提案手法の詳細を 2. 節で述べる．提案手法の評価を 4. 節で述べた後，5. 節で本稿のまとめについて述べる．

2.　b-彩色に基づくクラスタリング手法

2.1　表　　記

本稿では，太字のイタリック体大文字で集合を表記する．たとえば，$\boldsymbol{V}$ は頂点の集合を表す．また，$|\boldsymbol{V}|$ で集合 $\boldsymbol{V}$ の要素数を表記する．

表 1　非類似度行列

v_i	A	B	C	D	E	F	G	H	I
A	0								
B	0.20	0							
C	0.10	0.30	0						
D	0.10	0.20	0.25	0					
E	0.20	0.20	0.15	0.40	0				
F	0.20	0.20	0.20	0.25	0.65	0			
G	0.15	0.10	0.15	0.10	0.10	0.75	0		
H	0.10	0.20	0.10	0.10	0.05	0.05	0.05		
I	0.40	0.075	0.15	0.15	0.15	0.15	0.15	0.15	0

頂点集合 $\boldsymbol{V}$ と辺集合 $\boldsymbol{E} : \boldsymbol{V} \times \boldsymbol{V}$ に対し，$G(\boldsymbol{V}, \boldsymbol{E})$ でグラフを表記する．なお，以下では，任意の頂点間に多重辺やループがない単純グラフを考える．

また，$\forall v_i, v_j \in \boldsymbol{V}$ に対し，全ての頂点間の非類似度は関数 $d(v_i, d_j)$ で与えられると仮定し，関数 $d : \boldsymbol{V} \times \boldsymbol{V} \rightarrow R^+$ は対称性を満たすと仮定する (すなわち，$d(v_i, v_j) = d(v_j, v_i)$)．このため，本稿で考えるグラフ $G(\boldsymbol{V}, \boldsymbol{E})$ は無向グラフとなる．

1. 節で述べたように，クラスタリングを行うデータに対し，各頂点が各データに対応し，頂点を繋ぐ辺が頂点対間の非類似度のラベルを持つ辺重み付き完全無向グラフを考える．

［定義 1］（閾値グラフ）　非類似度行列 D と閾値 θ に対し，閾値グラフ $G(\boldsymbol{V}, \boldsymbol{E'})$ を以下で定義する．各データに頂点 v_i を対応付けた頂点集合 $\boldsymbol{V}$ に対応付け，全ての $\forall v_i, v_j \in \boldsymbol{V}$ に対し，$d(v_i, v_j) > \theta$ である場合に限って辺 (v_i, v_j) が $\boldsymbol{E}$ に含まれる．

2.2　b-彩色に基づくクラスタリング手法の概略

本節では文献 [3] で提案された b-彩色に基づくクラスタリング手法の概略について述べる．この手法では，各データは単一のクラスタ（色）に属すると仮定し，与えられた非類似度行列 D と閾値 θ に対して閾値グラフ $G(\boldsymbol{V}, \boldsymbol{E'})$ を構築し，以下の処理を通じて b-彩色性を満たすクラスタ（分割）を生成する．

1)　適正彩色を保ちつつ，最大の色数を持つように各頂点の色を決定する（この際には支配頂点は考慮しない）．

2)　支配頂点を持たない色を削除し，削除された色を持つ頂点の色を付け替える．

この手法では，上記 1), 2) ともに，欲張り法に基づいてバックトラックなしに実行する．1) においては，b-彩色では (i), (ii) の制約を満たす彩色の中で最大の色数を持つという性質を満たすために，適正彩色を満たしながら各頂点 v に隣接する頂点の色を相互に異なる色に彩色し，一旦彩色した色は固定する．2) においては，削除された色を持つ頂点 v の色を付け替える際，同じクラスタに属するデータ間の距離が極力近くなるように，頂点 v と最小の距離を持つ頂点の色に付け替える．

2.2.1　例

表 1 に示す非類似度行列 D を持つデータ集合が与えられたとする．図 1 は，閾値 θ=0.15 のもとでの表 1 に対する閾値グラフを示す．図 1 で，各辺は対応する非類似度のラベルが付け

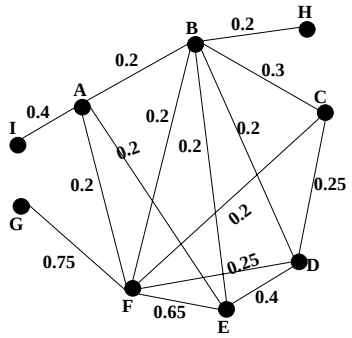

図 1　表 1 に示すデータに対する閾値 $\theta = 0.15$ での閾値グラフ

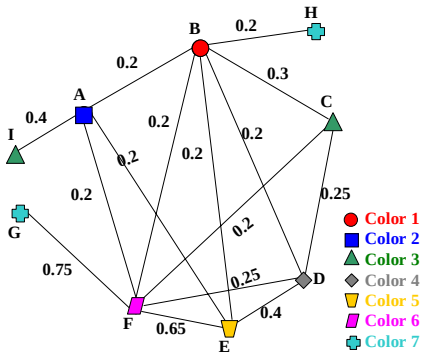

図 2　図 1 に対して手順 1) により適正彩色を保ちつつ最大の色数を持つように彩色されたグラフ

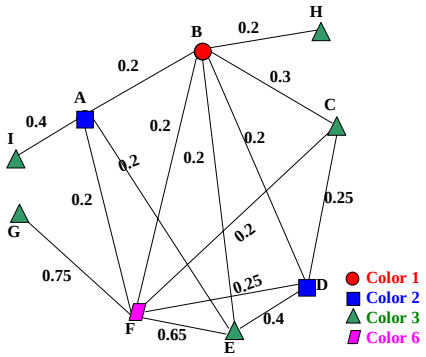

図 3　図 2 のグラフから手順 2) により支配頂点を持たない色を削除して色を付け替えたグラフ

られている．

図 1 の閾値グラフに対し，上記 1), 2) の処理を行うことにより，図 3 に示す分割（b-彩色グラフ）が構築される．図 3 で同じ色（形）を持つ頂点は同じクラスタに割り当てられる．このため，表 1 のデータはクラスタ {**A**,**D**}, {**B**}, {**C**,**E**,G,I}, {**F**} に分割される．ここで，各クラスタ内の太字の頂点は支配頂点となる．

同じ非類似度行列 D に対し，閾値 θ を変更することにより異なる閾値グラフが構築される．このため，異なる閾値に対して上記の手法を適用することにより，同じデータ集合に対しても異なる分割が得られる．

2.3　クラスタリングの妥当性指標

これまで，クラスタリング手法を適用して得られる分割に対する妥当性を測るため，多くの指標が提案されている [1], [5]．これらの指標のうち，本稿では一般化 Dunn 指標（$Dunn_G$ と表記）に焦点を当てる．この指標は，クラスタ間分離性とクラスタ内凝集性のバランスを考慮した指標であり，コンパクトで分離性の良いクラスタへの分割を考える上で適した指標と考えられる．

頂点集合$\boldsymbol{V}$が分割 $\boldsymbol{P} = \{C_1, C_2, \ldots, C_k\}$ で表現されるクラスタに分割されたとする．ここで，$\forall C_i, C_j \in \boldsymbol{P}, C_i \cap C_j = \phi \ for \ i \neq j$. が成り立つと仮定する．b-彩色に基づくクラスタリングにおいては，各クラスタ $C_i \in \boldsymbol{P}$ は 1 つの色に対応し，異なるクラスタが同じ色に対応することはないため，以下では，$\boldsymbol{P}$ という記法で，クラスタの集合と色の集合を表すこととする．

［定義 2］（クラスタ内平均非類似度）　全ての $\forall C_h \in \boldsymbol{P}$ に対し，クラスタ内平均非類似度 $S_a(C_h)$ を以下で定義する．

$$S_a(C_h) = \frac{1}{\eta_h(\eta_h - 1)} \sum_{o=1}^{\eta_h} \sum_{o'=1}^{\eta_h} d(v_o, v_{o'}) \tag{1}$$

ただし，$\eta_h = |C_h|$，$v_o, v_{o'} \in C_h$ （C_h は頂点集合を含むため，$|C_h|$ でその要素数を表す）．

［定義 3］（クラスタ間平均非類似度）　全ての $\forall C_i, C_j \in \boldsymbol{P}$ に対し，クラスタ間平均非類似度を以下で定義する．

$$d_a(C_i, C_j) = \frac{1}{\eta_i \eta_j} \sum_{p=1}^{\eta_i} \sum_{q=1}^{\eta_j} d(v_p, v_q) \tag{2}$$

ただし，$\eta_i = |C_i|$，$\eta_j = |C_j|$，$v_p \in C_i$, $v_q \in C_j$.

［定義 4］（一般化 Dunn 指標）　分割$\boldsymbol{P}$に対する一般化 Dunn 指標 (以下，$Dunn_G$ と表記) を以下で定義する．

$$Dunn_G(\boldsymbol{P}) = \frac{\min_{i,j,i \neq j} d_a(C_i, C_j)}{\max_h S_a(C_h)} \tag{3}$$

ただし，$C_h, C_i, C_j \in \boldsymbol{P}$.

一般化 Dunn 指標は，分割$\boldsymbol{P}$におけるクラスタ間の分離性に対する指標として提案された Dunn 指標がアウトライアーなどに対する頑健性に欠けるという問題に対し，指標の計算に全てのデータを陽に取り込むことで頑健性を高めた指標である．基本的には，与えられたデータに対して $Dunn_G(\boldsymbol{P})$ が大きな分割を与えるクラスタリング手法ほど，アウトライアーに相当するデータが含まれた場合にも頑健なクラスタを生成するという意味で良い手法と考えることができる．

文献 [3] での手法では，同じデータに対しても閾値 θ をかえることで異なる分割が得られる．このため，表 1 に示すデータに対しても，閾値をかえることで異なる $Dunn_G$ を持つ分割が得られる．この例では，閾値を変えて適用することにより，θ=0.15 において最大の $Dunn_G$ (1.522) が得られるため，表 1 のデータに対して文献 [3] の手法を適用して得られる最良の結果は図 3 となる．

3.　b-彩色に基づくクラスタリングの色付け替え手法

3.1　b-彩色に基づくクラスタリングの更新

2.2 節で述べたように，表 1 のデータに対して [3] の手法を

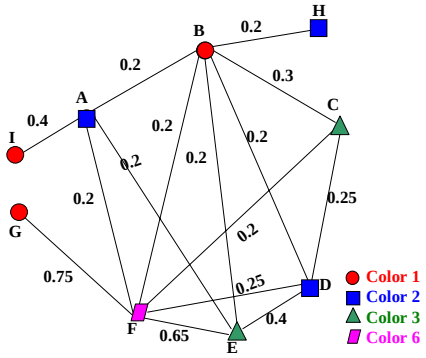

図 4　Another b-coloring with larger $Dunn_G$ for Table 1.

適用すると，図 3 に示す分割が $Dunn_G$ の観点からは最良の結果として得られる．

しかし，表 1 のデータに対しては，同じ色数を持つ分割であっても，図 1 に示す b-彩色グラフより大きな $Dunn_G$ を持つ別の b-彩色グラフが存在する．その一例を図 4. に示す．容易に確認できるように，図 4 に示すグラフは b-彩色グラフであり，しかも $Dunn_G = 1.538$ である．このため，一般化 Dunn 指標の観点からは，図 3 よりもより良いクラスタリング結果と考えられる．

この例で示すように，一般には文献 [3] の手法で得られる分割図 1 に示すグラフに対してはより大きな $Dunn_G$ を持つ別の b-彩色グラフが存在する．このため，b-彩色性を満たしながらクラスタリング指標の観点から（本稿では $Dunn_G$ を用いる）より良い分割を得ることが重要になる．この問題を以下のように定式化する．

［問題 1］（b-彩色クラスタリングにおける色付け替え問題）　グラフ G に対して b-彩色性を満たす分割$\boldsymbol{P}$が与えられた際，b-彩色性を満たしながら分割$\boldsymbol{P}$よりも良い分割$\boldsymbol{P'}$を見つけよ

上記で述べたように，本稿では分割$\boldsymbol{P}$の良さの指標として $Dunn_G(\boldsymbol{P})$ を用いる．以下では，問題 1 に対する本稿での提案手法を述べる．

3.2　定　　義

各頂点 $v \in \boldsymbol{V}$ に対し，関数 $N(v)$ は頂点 v に隣接する頂点の集合を返す．すなわち，全ての $\forall v' \in N(v)$ に対し，頂点対 (v, v') は辺集合$\boldsymbol{E}$ に含まれる．

関数 $c(v)$ は頂点 v の色を返し，関数 $N_c(v)$ は頂点 v に隣接する頂点集合に対する色集合を返す．(すなわち，$N_c(v) = \cup_{v' \in N(v)} c(v')$).

関数 $C_p(v)$ を$\boldsymbol{P} \setminus N_c(v)$ として定義する．(ここで，$\boldsymbol{P}$は分割$\boldsymbol{P}$における全ての色集合を表す)．なお，$C_p(v)$ は頂点 v に付けられていた元の色 $c(v)$ も含むことに注意されたい．

3.2.1　Critical 頂点集合と non-critical 頂点集合

集合$\boldsymbol{V_d}$でグラフ G における支配頂点の集合を表す．各頂点 $v_d \in \boldsymbol{V_d}$ に対し，もし頂点 $v_s \in N(v_d)$ が集合 $N(v_d)$ において色 $c(v_s)$ を持つ唯一の頂点である場合，頂点 v_s を頂点 v_d の支持頂点と呼ぶ．

頂点集合$\boldsymbol{V}$を，$\boldsymbol{V_c} \cap \boldsymbol{V_{nc}} = \phi$ であるような直和 $\boldsymbol{V_c} \sqcup \boldsymbol{V_{nc}}$ に分解する．集合 $\boldsymbol{V_c}$ に含まれる各頂点 $v_c \in \boldsymbol{V_c}$ を critical 頂点と呼び，また，集合 $\boldsymbol{V_{nc}}$ に含まれる各頂点 $v_{nc} \in \boldsymbol{V_{nc}}$ を non-critical 頂点と呼ぶ．本稿で提案する手法は，集合 $\boldsymbol{V_{nc}}$ に含まれる頂点 $v_{nc} \in \boldsymbol{V_{nc}}$ の色付け替えを試みる．他方，集合 $\boldsymbol{V_c}$ に含まれる頂点 $v_c \in \boldsymbol{V_c}$ の色の変更は行わない．

集合$\boldsymbol{V_c}$を，更に，重なりのない直和 $\boldsymbol{V_d} \sqcup \boldsymbol{V_s} \sqcup \boldsymbol{V_f}$ に分解する．ここで，集合 $\boldsymbol{V_f}$ は色付け替えの試みが完了した頂点の集合を保持する．

頂点 $v \in \boldsymbol{V}$ の色 $c(v)$ を c に色付け替えした際，他の頂点が支配頂点になったり支持頂点になることにより，ある頂点 $v_{nc} \in \boldsymbol{V_{nc}}$ が新しく critical 頂点になる可能性がある．頂点 v の色付け替えに伴うグラフ G における色の変化を反映するため，集合 $\boldsymbol{V_c^{tmp}}(v, c)$ で色付け替えにより新しく critical 頂点になった頂点の集合を表し，$\boldsymbol{P}(v, c)$ でその際の分割を表現することとする．分割 $\boldsymbol{P}(v, c)$ においては，頂点 v の色 $c(v)$ のみが c に付け替えられ，頂点 v 以外の他の頂点 $\forall v' \in \boldsymbol{V}$, $v' \neq v$ の色は変更されない．

［定義 5］（頂点とクラスタの平均非類似度）　全ての $\forall v \in \boldsymbol{V}$, $\forall C_i \in \boldsymbol{P}$ に対し，頂点 v とクラスタ C_i の平均非類似度を以下で定義する．

$$d_a(v, C_i) = \frac{1}{\eta_i} \sum_{p=1}^{\eta_i} d(v, v_p) \tag{4}$$

ここで，$\eta_i = |C_i|$, $v_p \in C_i$.

3.3　b-彩色に基づくクラスタリングの色付け替え手法

3.3.1　色付け替えのアプローチ

支配頂点の定義より，頂点 $v_d \in \boldsymbol{V_d}$ は他の全ての色の頂点と接続しており，頂点 v_d は他のクラスタから少なくとも閾値 θ 以上は離れており，クラスタ間非類似度の大きな分割を生成するのに貢献する．このため，本稿のアプローチでは，良い分割を構成する上で頂点 v_d は重要であり，その色は変化すべきでないと考える．

また，集合$\boldsymbol{V_s}$に含まれる頂点 v_s の色 $c(v_s)$ を変更すると，集合$\boldsymbol{V_d}$に含まれるある支配頂点 v_d が他の全ての色に接続しなくなる恐れがある．このため，色 $c(v_s)$ の変更は分割の質を低下させる可能性があるため，本稿では集合$\boldsymbol{V_s}$に含まれる頂点 v_s の色も変更しない．

上記により，本稿では集合 $\boldsymbol{V} \setminus \{\boldsymbol{V_d} \sqcup \boldsymbol{V_s}\}$ に含まれる頂点 v の色の変更を考える．さらに，色付け替えの停止性を保証するため，各頂点の色付け替えを一度のみ試みる．これを実現するため，頂点 $v \in \boldsymbol{V}$ の色付け替えを試した後，頂点 v を集合 $\boldsymbol{V_f}$に移動する．

まとめると，本稿では集合 $\boldsymbol{V} \setminus \{\boldsymbol{V_d} \sqcup \boldsymbol{V_s} \sqcup \boldsymbol{V_f}\} = \boldsymbol{V_{nc}}$, に含まれる頂点 v の色付け替えを考え，頂点 v の色付け替えを試みた後，v を集合$\boldsymbol{V_f}$に移動することでその色を固定する．このため，集合$\boldsymbol{V_{nc}}$の要素数は色付け替えにより単調に減少する．

また，集合 $\boldsymbol{V_{nc}}$ に含まれる頂点 v_{nc} の色 $c(v_{nc})$ は，色付け替えに伴い変更される可能性がある．このため，集合$\boldsymbol{V_{nc}}$に含まれる頂点を分割$\boldsymbol{P}$の質の評価に用いると，その評価の信頼性が損なわれる恐れがある．このため，分割$\boldsymbol{P}$の質の評価には集

合$\boldsymbol{V_{nc}}$に含まれる頂点は考慮せず，あくまで色が確定した集合$\boldsymbol{V_c}$に含まれる頂点のみを考慮することとする．

3.3.2 頂点選択規範

集合 $\boldsymbol{V_{nc}}$ に含まれる頂点 v_{nc} の中から，本稿では $d_a(v, c(v))$ が最大となる頂点 $v \in \boldsymbol{V_{nc}}$ を色付け替え頂点として選択する．すなわち，$v^* = \arg\max_{v \in \boldsymbol{V_{nc}}} d_a(v, c(v))$ を選択規範とする．ここで，$d_a(v, c(v))$ は各頂点 v とそれが割り当てられたクラスタ $c(v)$ とのクラスタ内平均非類似度であるため，頂点 v が割り当てられたクラスタ（色）$c(v)$ にとってどのくらい外れ値であるかに対する指標とみなすことができる．このため，より外れ値の度合いが大きな頂点から色付け替えを試みる．

他方，上記で最大ではないような頂点 v'，すなわち $v' \neq \arg\max_{v \in \boldsymbol{V_{nc}}} d_a(v, c(v))$ であるような頂点 v' が頂点 v^* より先に色付け替えを試みられたと仮定する．この場合，頂点 v' の色付け替えに伴い集合 $\boldsymbol{V_{nc}}$ に含まれる頂点 v'' が集合 $\boldsymbol{V_c}$ に移動する可能性があるため，頂点 v^* を先に色付け替えを試みた場合と比較すると，頂点 v^* に隣接する色の集合 $N_c(v^*)$ が増加する可能性がある．頂点 v^* が取り得る色により制約が付加されるため，頂点 v^* の色の候補数 $|C_p(v^*)|$ が減少する可能性がある．このため，もし頂点 v' が頂点 v^* より先に色付け替えを試みられたすると，式 (4) の分子を減少させることにより $Dunn_G$ を最大化するような色とは異なった色に頂点 v^* の色が付け替えられてしまう恐れがある．

上記の理由により，頂点の選択規範として，頂点とクラスタの平均非類似度 $d_a(v, c(v))$ 最大化規範に基づき色付け替え頂点を選択する．

3.3.3 色選択規範

本稿の手法では，適正彩色を保ちつつ色を付け替えることにより分割$\boldsymbol{P}$の質を向上させることを試みる． 3.3.2 節で述べた規範に基づき頂点 v を選択した場合，頂点 v の色付け替え候補となる集合 $C_p(v)$ に含まれる色を調べ，式 (3) で定義される $Dunn_G$ を最大化する色を選択する．

頂点 v の色付け替えに伴い，$Dunn_G$ の計算に必要な $S_a(\cdot)$ と $d_a(\cdot, \cdot)$ の値も変化する．このため，頂点 v の $c(v)$ が変更された場合には，これらの値も再計算して更新する必要がある．大規模なデータに対応するためには，色付け替えに伴う更新に要する計算量を抑え効率的な再計算を実現する必要がある．本稿では，それぞれの値を保存しておき，保存した値を活用して効率的な再計算を実現する．特に，単純な $d_a(\cdot, \cdot)$ の再計算は $O(n^2)$ 要するが，保存した値を用いることで $O(n)$ に抑えることが可能となる．

$\boldsymbol{P}$ に含まれる $\exists\, C$,C_i に対し，クラスタ C に割り当てられていた頂点 $v \in \boldsymbol{V_{nc}}$ がクラスタ C_i に割り当てられた（色が付け替えられた）とする．これに伴い，頂点 v は集合 $\boldsymbol{V_f}$ に移動される．分割 $\boldsymbol{P} = \boldsymbol{P} \backslash \{C, C_i\} \sqcup \{C, C_i\}$, に対し，保存しておいた値を用いることにより，定義 2 のクラスタ内平均非類似度に対する更新後の値を下記により再計算する．

Algorithm 1 A Greedy Re-coloring Algorithm for b-coloring based Clustering

Require: $G(\boldsymbol{V}, \boldsymbol{E})$; *//A graph with a set of vertices and a set of edges.*

Require: $\boldsymbol{P}$; *//a partition which is a b-coloring of* $G(\boldsymbol{V}, \boldsymbol{E})$

1: $\boldsymbol{C'} = \boldsymbol{P}$;
2: Divide $\boldsymbol{V}$ into $\boldsymbol{V_c} \sqcup \boldsymbol{V_{nc}}$
3: **while** $\boldsymbol{V_{nc}} \neq \phi$ **do**
4: $\quad v^* = \arg\max_{v' \in \boldsymbol{V_{nc}}} d_a(v', c(v'))$; *//vertex selection*
5: $\quad$ **for** each $c \in C_p(v^*)$ **do**
6: $\quad\quad$ create $\boldsymbol{V_c^{tmp}}(v^*, c)$ and $\boldsymbol{P}(v^*, c)$ induced from the re-coloring of $c(v^*)$ into c;
7: $\quad\quad$ For $\forall C_i \in \boldsymbol{P}(v^*, c)$, calculate $d_a(v^*, C_i)$ w.r.t $\boldsymbol{V_c} \sqcup \boldsymbol{V_c^{tmp}}(v^*, c)$;
8: $\quad\quad$ calculate $S_a^{new}(C_h)$ and $d_a^{new}(C_i, C_j)$ for $\forall C_h, C_i, C_j \in \boldsymbol{P}(v^*, c)$;
9: $\quad$ **end for**
10: $\quad c^*(v) = \arg\max_{c \in C_p(v)} Dunn_G(\boldsymbol{P}(v, c))$;
11: $\quad$ Re-color $c(v*)$ to c^* in $\boldsymbol{C'}$; *//re-coloring of* $\boldsymbol{C'}$
12: $\quad \boldsymbol{V_{nc}} = \boldsymbol{V_{nc}} \backslash \{v\}$;
13: $\quad \boldsymbol{V_f} = \boldsymbol{V_f} \cup \{v\}$;
14: $\quad \boldsymbol{V_c} = \boldsymbol{V_c} \cup \boldsymbol{V_c^{tmp}}(v, c)$ *//*$v \in \boldsymbol{V_c^{tmp}}(v, c)$ *into* $\boldsymbol{V_d}$ *or* $\boldsymbol{V_s}$ *or* $\boldsymbol{V_f}$ *due to its property*
15: **end while**
16: **return** $\boldsymbol{C'}$;

$$S_a^{new}(C) = S_a^{old}(C) \tag{5}$$

$$S_a^{new}(C_i) = \frac{|C_i| S_a^{old}(C_i) + |C| d_a(v, C_i)}{|C_i| + 1} \tag{6}$$

$$S_a^{new}(C_j) = S_a^{old}(C_j) \quad \forall C_j \in \boldsymbol{P} \backslash \{C, C_i\} \tag{7}$$

同様に，定義 2 のクラスタ間平均非類似度を，下記の更新式を用いて計算する．

$$d_a^{new}(C, C_i) = \frac{|C||C_i| d_a^{old}(C, C_i) + |C| d_a(v, C)}{|C|(|C_i| + 1)} \tag{8}$$

$$d_a^{new}(C_i, C_j) = \frac{|C_i||C_j| d_a^{old}(C_i, C_j) + |C_j| d_a(v, C_j)}{(|C_i| + 1)||C_j|} \tag{9}$$

$$d_a^{new}(C_j, C_h) = d_a^{old}(C_j, C_h) \quad \forall C_j, C_h \in \boldsymbol{P} \backslash \{C, C_i\} \tag{10}$$

3.3.4 色付け替え手法アルゴリズム

提案する色付け替えアルゴリズムをアルゴリズム 1 に示す．4 行目と 10 行目で同じ値を持つ候補が複数ある場合にはランダムに選択することとしている．

提案するアルゴリズムはクラスタリングに対して下記の性質を持つ．

［命題 1］（適正彩色性） アルゴリズム 1 は，分割$\boldsymbol{P}$から適正彩色性を満たす分割$\boldsymbol{P'}$を生成する．

証明 アルゴリズム 1 は，5 行目以下で頂点集合$\boldsymbol{V}$に含まれる各頂点 v の色 $c(v)$ を $C_p(v)$ に含まれる色 c' にのみ付け替える．$C_p(v)$ の定義より，$C_p(v)$ に含まれる全ての色 c' に対し，c' は $N_c(v)$ に含まれない．このため，適正彩色性が保証される．□

［命題 2］（b-彩色性） アルゴリズム 1 は分割$\boldsymbol{P}$から b-彩色性

を満たす分割$\boldsymbol{P'}$を生成する．

証明　命題 1 よりアルゴリズム 1 が生成する分割$\boldsymbol{P'}$は適正彩色性を満たす．このため，b-彩色性を示すためには，各色（クラスタ）に対して少なくとも 1 つの支配頂点が存在することを示せばよい．定義より，分割$\boldsymbol{P}$においては各色（クラスタ）に対して少なくとも 1 つの支配頂点が存在する．アルゴリズム 1 は集合$\boldsymbol{V_d}$に含まれる任意の頂点 v_d の色を変更しないため，分割$\boldsymbol{P'}$においては各色（クラスタ）に対して少なくとも 1 つの支配頂点が存在する．□

［命題 3］（$Dunn_G(\boldsymbol{P})$ の単調性）　グラフ $G(\boldsymbol{V},\boldsymbol{E})$ に対する分割$\boldsymbol{P}$に対し，アルゴリズム 1 が生成する分割$\boldsymbol{P'}$では $Dunn_G(\boldsymbol{P'})$ が単調に増加する．

証明　アルゴリズム 1 での 10 行目と 11 行目において，$Dunn_G(\boldsymbol{P})$ を最大化する色 $c(v^*)$ が選択される (色が変更されない場合，すなわち，$c_{new}(v)$=c(v) もあり得ることに注意）．この処理が集合 $\boldsymbol{V_{nc}}$ に含まれる全ての頂点 v_{nc} に対して実行されるため，アルゴリズム 1 が停止した際には $Dunn_G(\boldsymbol{P'})$ は単調に増加する．□

4. 評　　価

3. 節で提案した手法を，UCI リポジトリ [7] で提供される *Zoo* と *Mushroom* という 2 つのベンチマークデータに対して評価した．提案した手法（以下では「色付け替え彩色法」と表記）で得られる結果（分割）を，以下の 3 手法と比較した．

b-彩色法　文献 [3] での手法を用いて $Dunn_G$ を最大化した分割

Hansen 法　最小彩色に基づく手法 [6] で得られる分割

最短距離法　併合法に基づく最短距離法 [9] で得られる分割

文献 [10] にならい，定義 4 での一般化 Dunn 指標に加え，識別性 *(Distinctness)* [10] と呼ばれる指標での評価も行った．この指標はクラスタ数やデータ間の非類似度とは独立した指標であるため，指定するクラスタ数や使用するデータ間の非類似度が異なった場合に対しても汎用的な指標である．

［定義 6］（分布適合性分散）　分割$\boldsymbol{P}$におけるクラスタ C_k と C_l の分布適合性を以下で定義する．

$$Var(C_k,C_l) = \frac{1}{m}\sum_{i}^{m}\sum_{j}\left(P\left(a_i=V_{ij}|C_k\right)-P\left(a_i=V_{ij}|C_l\right)\right)^2 \quad (11)$$

ここで，m はデータを特徴付ける属性 a_i の数であり，$P(a_i = V_{ij}|C_k)$ はクラスタ C_k において属性 a_i が属性値 V_{ij} を取る条件付確率である．

定義 6 ではデータは各属性に対して単一の値を持つと仮定している．クラスタ C_k と C_l が「離れて」いるほど，各クラスタにおける条件付き確率の値が異なる．このため，$Var(C_k,C_l)$ が大きいほどクラスタ C_k と C_l の非類似度が大きいと解釈できるため，C_k と C_l 間の分離の程度を表す指標である．

［定義 7］（識別性 (Distinctness)）　クラスタ $\{C_1,C_2,..,C_p\}$ で構成される分割$\boldsymbol{P}$に対し，識別性 *(Distinctness)* を分割$\boldsymbol{P}$における分布適合性分散の平均として定義する．

表 2　Zoo データセットに対する評価

手法	クラスタ数	識別性	$Dunn_G$
色付け替え b-彩色法	7	0.652	1.120
b-彩色法	7	0.612	1.071
最短距離法	2	0.506	0.852
Hansen 法	4	0.547	1.028

表 3　Evaluation of Mushroom Data.

手法	クラスタ数	識別性	$Dunn_G$
o 色付け替え b-彩色法	17	0.728	0.995
b-彩色法	17	0.713	0.891
最短距離法	20	0.615	0.866
Hansen 法	19	0.677	0.911

$$\text{識別性} = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} Var(C_k,C_l)}{p\times(p-1)} \quad (12)$$

２つのクラスタ間の分離の程度を表す分布適合性分散の平均として定義されるため，分割を比較する際には，より大きな識別性を持つ分割の方が各クラスタが異なる概念を表現することになり，良い分割と見なすことができる [10].

4.1　Zoo データセットに対する評価

Zoo データセットは，17 個の属性と 7 個のクラスで表現される 100 個の事例から成る [7]．最初の属性は動物名を表し，15 個の属性はブーリアン属性であり，最後の属性は肢の数を表す数値属性である．

評価結果を表 2 に示す．表 2 より，本稿で提案した色付け替え b-彩色法が識別性に対して最良の結果を得たことがわかる．この結果より，支配頂点を考慮した色付け替えが分離性の高い分割を得ることに効果的であることがわかる．他方，色付け替え b-彩色法は $Dunn_G$ に対しても最良の結果を得ており，提案した色付け替え法がもとの b-彩色法で得られる分割を向上させることに役立つことがわかる．

4.2　Mushroom データセットに対する評価

Mushroom データセットでは，各データは 1 つのきのこに対応し，色，臭い，サイズ，形などの 21 個の物理的性質が記述され，毒性持つか食用に適するかがクラスラベルとして記述される．なお，Zoo データセットと異なり，全ての属性は名義属性である．また，8124 個のデータが含まれ，Zoo データセットと比較して大きなサイズとなっている．

評価結果を表 3 に示す．Zoo データセットに対する結果と同様に，本稿で提案した色付け替え b-彩色法は，識別性の観点からも $Dunn_G$ の観点からも最良の結果となっており，特に [3] での手法よりもよい結果を得ていることが確認できる．

5. お わ り に

本稿では，b-彩色性を保ちながらクラスタ分割の質を向上させる色付け替え手法を提案した．b-彩色に基づくクラスタリング手法 [3] はクラスタ数が指定されない場合でも精緻な分割を与えることができ，クラスタリング手法としても様々な特徴を

備える．しかし，グラフの b-彩色によりクラスタリングを行う際，クラスタの質は明示的に考慮されていなかった．本稿で提案した手法は，この点を補完する役割を果たし，b-彩色という性質を保証しながらクラスタの質の向上を実現する．提案した手法を実装し，UCI リポジトリのベンチマークデータに対する評価を通じて提案手法の有効性を確認した．特に，一般化 Dunn 指標と識別性に対する比較を通じて，提案手法はクラスタリングにおいて重要となるクラスタ間分離性とクラスタ内凝集性のバランスをほどよく実現することを確認した．

謝　　辞

本研究を支援してくださいました田中譲教授に謝意を表します．本研究の一部は，日本学術振興会先端研究拠点事業–拠点形成型– (No.18001) および文部科学省科研費若手研究 (B) (No.18700131) の補助による．

文　　献

[1] J.C. Bezdek and N.R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics*, 28(3):301–315, 1998.

[2] B. Effantin and H. Kheddouci. The b-chromatic number of some power graphs. *Discrete Mathematics and Theoretical Computer Science*, 6(1):45–54, 2003.

[3] H. Elghazel, V. Deslandres, M.S. Hacid, A. Dussauchoy, and H. Kheddouci. A new clustering approach for symbolic data and its validation: Application to the healthcare data. In F.Esposito et al.(Eds), editor, *ISMIS2006 (Springer Verlag LNAI 4208)*, pages 473–482, 2006.

[4] A. Guënoche, P. Hansen, and B. Jaumard. Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *Journal of Classification*, 8:5–30, 1991.

[5] S. Günter and H. Bunke. Validation indices for graph clustering. *Pattern Recognition Letters*, 24:1107–1113, 2003.

[6] P. Hansen and M. Delattre. Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association*, 73:397–403, 1978.

[7] S. Hettich, C.L. Blake, and C.J. Merz. Uci repository of machine learning databases, 1998.

[8] W. Irving and D. F. Manlov. The b-chromatic number of a graph. *Discrete Applied Mathematics*, 91:127–141, 1999.

[9] A.K. Jain, M.N. Murty, and FlynnT. P.J. Data clustering: A review. *ACM Computing Surveys*, 31:264–323, 1999.

[10] M. Kalyani and M. Sushmita. Clustering and its validation in a symbolic framework. *Pattern Recognition Letters*, 24(14):2367–2376, 2003.