クラスタリング結果の特徴抽出を用いる高次元データの対話的クラスタリング

中村朋健[†] 上土井陽子[††]
若林真一[††] 吉田典可[††]

クラスタリング手法によって抽出されたクラスタがユーザにとって役立つ結果であるかどうかを判断することは難しい問題である．我々はユーザの要求に近いクラスタを対話的に導くことを目標とした対話的クラスタリング手法（ICM: Interactive Clustering Method）とその構成要素である特徴抽出手法（FEM: Feature Extraction Method）および選択指標関数（IF: Indicator Function）を提案する．対話的クラスタリング手法は，ユーザが自身の要求に応じたクラスタリング結果を出力するための手法であり，クラスタリング手法の入力パラメータ設定を支援することを目標とする．特徴抽出手法はクラスタ集合から，各属性に関して特異な分布を持つクラスタを特徴として抽出する手法であり，得られた特徴をユーザに提示する．選択指標関数は特徴抽出手法により得られた特徴を入力して，2 つのクラスタリング結果において，どちらが自然なクラスタリング結果であるかを示す関数である．本稿では，ユーザが対話的クラスタリングによって提供される情報を基に自身の要求に応じたクラスタリング結果を得ようとする過程を示すことで対話的クラスタリング手法の有効性を示す．

Interactive Clustering Based on a Feature Extraction Method for Clustering Results on High Dimensional Data Sets

Tomotake Nakamura,[†] Yoko Kamidoi,[††]
Shin'ichi Wakabayashi[††] and Noriyoshi Yoshida[††]

It is difficult to judge whether clustering results obtained by a clustering method are useful for users. In order to bridge a gap between user's needs and a user's level of technique for controlling of clustering algorithms, we propose the Interactive Clustering Method (ICM), the Feature Extraction Method (FEM) and the Indicator Function (IF), where FEM and IF are parts of ICM. ICM consists of a user, a clustering method, a method which outputs a relation between two clustering results, FEM which can extract a clustering result, and a function which outputs an indicator for selecting a desired clustering result. FEM can extract clusters which differ from other clusters and distribution of the clusters as a feature of a clustering result and output the features in a form as available to the user. By using IF, users can choose a more natural clustering result. In this paper, we focus on ICM's application to decision supports on user's setting of input parameters of a clustering algorithm until user's needs are satisfied. In simulation experiments, we show the process of acquiring a desired clustering result by ICM and we show effectiveness of ICM.

1. まえがき

1.1 背景・目的

近年，情報化社会となり，また記憶装置の低価格化が進むことによりデータセットに蓄積されるデータは高次元であり大規模なものとなっている．今後はさらに高次元かつ大規模なデータセットが構築されることが予想される．高次元かつ大規模なデータセットからユーザにとって有用な情報を効率良く抽出するために，多くのデータマイニングツールが開発されている[5),10),13)]．多くのデータマイニングアルゴリズムの主要目的は大規模データから簡潔，かつ，説明可能な情報の発見を手助けすることである．しかし，現在の多くのデータマイニングアルゴリズムはその目的に達していない[2)]．データマイニングの 1 つの技法であるクラスタリングを利用すれば，自然なクラスタ，つま

† 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University
†† 広島市立大学情報科学部
Faculty of Information Sciences, Hiroshima City University

り類似したデータ要素の集合を見つけ出せる．一般的なクラスタリング手法は，入力パラメータを変化させることで，多くの自然なクラスタリング結果を出力できる．しかし，得られたクラスタリング結果が要求に近いクラスタ集合を形成しているかどうかをユーザが判断することは難しく，クラスタリング手法を効率良く利用できないことがある．ある入力パラメータでの結果がユーザの要求に近いクラスタリング結果であるのかどうかをユーザが容易に判断できる情報の提供が望まれる．

クラスタリング手法は類似検索，顧客分割，パターン認識，動向分析，そしてクラシフィケーションなどの幅広い分野で使用される．これらの分野では各ユーザによって欲する解が異なることが多いため，我々はクラスタリング問題の結果の導出を自動化するのではなく，対話的に各ユーザの欲する結果を導くことを目標とする．文献 3) では，ユーザとコンピュータが協力してベストなクラスタが現れる部分空間を調べるアルゴリズムが提案されている．データセットからユーザの欲するクラスタを見つけ出すことは難しい問題であるため，このアルゴリズムのようにユーザとコンピュータが対話的に情報をやりとりしてユーザの要求に近いクラスタ集合を見つけ出す必要がある．特に高次元なデータセットが入力であるとき，自然なクラスタリング結果が多く存在し，かつ，ユーザがクラスタリング結果の性質を理解し難いので，ユーザの要求と現在得られているクラスタリング結果の性質が近いかどうかをユーザが理解することは難しい問題である．我々の目標は，高次元な入力データ空間のままでパラメータを変化させることにより，ユーザの要求に近いクラスタリング結果を効果的に導くことができるように支援することである．

2 つのクラスタリング結果を用いてより自然な形を算出することができるならば，ユーザは自身の要求に近い結果を得やすいであろう．実際，高次元データの結果の解釈や，結果が要求にどの程度沿うものかを判断することは容易ではない．我々は，クラスタリング結果の特徴的な情報を自動的に抽出することでユーザの高度な判断を支援し，クラスタリング結果やその特徴をユーザが詳細に解析することなく，2 つのクラスタリング結果から，より自然なクラスタリング結果を選択可能にしたいと考えている．我々はクラスタリング結果の特徴的な情報を自動的に抽出するために特徴抽出手法を提案し，2 つのクラスタリング結果から，より自然なクラスタリング結果を選択可能にするために選択指標関数を提案する．特徴抽出手法と選択指標関数は，本稿で我々が提案する対話的にユーザが自身の要求に近いクラスタリング結果を得ることが可能な対話的クラスタリング手法の構成要素である．

1.2 対話的クラスタリングの概要と本稿の構成

対話的クラスタリング手法にはユーザ，クラスタリングアルゴリズム，特徴抽出手法，選択指標関数，そして 2 つのクラスタリング結果の関係を調べる手法が含まれる．対話的クラスタリング手法への入力はユーザがクラスタリングしたいデータセットであり，出力はユーザの要求に近いクラスタ集合である．対話的クラスタリング手法内ではユーザの要求に近いクラスタ集合が抽出されるまで繰り返しユーザとクラスタリング手法間で情報交換が行われる．対話的クラスタリングでは，ユーザにクラスタリング結果，クラスタリング結果の特徴，2 つのクラスタリング結果におけるクラスタ間のデータ要素の交差関係，そして 2 つのクラスタリング結果の選択指標の情報を与える．特徴抽出手法はクラスタリング結果の特徴を抽出する手法であり，任意のクラスタリング手法によって出力されたクラスタ集合を入力とし，各属性において特異な分布を持つクラスタとその分布を出力する．選択指標関数は特徴抽出手法による特徴を入力として，2 つのクラスタリング結果からより自然なクラスタリング結果がどちらのクラスタリング結果であるかを判定する指標をユーザに提供する．ユーザは選択指標に基づいたクラスタリング結果の選択をする必要はなく，最終的なクラスタリング結果の選択はユーザに委ねるものとする．

本稿では，2 章において，対話的にクラスタリングするための手法である対話的クラスタリング手法とその構成要素である特徴抽出手法および選択指標関数を提案する．3 章において，合成 2 次元入力データを用いてユーザの意思決定支援の具体例を示す．4 章において，2 クラス分類問題に対応するベンチマークデータを用いたシミュレーション実験により，本稿で提案する対話的クラスタリング手法を実験的に評価する．4 章では，まず，ユーザが 2 つのクラスタリング結果を選択するときの支援となる選択指標の有効性を評価する．次に，選択指標に基づいて入力パラメータ値を設定することで得られたクラスタリング結果をクラスタリング向け分類エラーの観点から評価する．5 章で，本稿をまとめる．

2. 対話的クラスタリング手法（ICM）

本稿で提案する対話的クラスタリング手法（ICM: Interactive Clustering Method）を説明するために基本的な語句や変数を 2.1 節で定義する．2.2 節では対

話的クラスタリングアルゴリズムを提案する．2.3 節ではクラスタリング結果の特徴を抽出する特徴抽出手法（FEM: Feature Extraction Method）を提案し，2.4 節では 2 つのクラスタリング結果から一方より自然なクラスタリング結果を示すことが可能な選択指標関数（IF: Indicator Function）を提案する．

2.1 定　　義

2.2 節以降において，我々が提案するアルゴリズムを説明するために，基本的な語句や変数などの表記を定義する．ユーザがクラスタリングしたいと考えているデータセットをオリジナルデータセット ODS と呼ぶ．オリジナルデータセット ODS の属性数を dim とする．クラスタリング手法によってユーザの要求する出力を得やすくするために，オリジナルデータセット ODS は変換関数（TF: Transform Function）によってデータセット DS に変換される．データセット DS はすべて数値データであるとする．ユーザがクラスタリング手法（CM: Clustering Method）に与える入力パラメータ値の集合を IP_{CM} とする．クラスタリング手法 CM への入力は IP_{CM} とデータセット DS であり，クラスタリング手法 CM の出力はクラスタリング結果 $\mathcal{C} = \{C_1, \ldots, C_{N_\mathcal{C}}, Noise_\mathcal{C}\}$ であるとする．ここで，C_i（$1 \leq i \leq N_\mathcal{C}$）はデータセット DS の部分集合であり，類似したデータ要素の集合と見なされた 1 つのクラスタである．また，データセット DS の一般的な性質に適合しないデータ要素，または，異なる機構により作成されたことが疑われるデータ要素を外れ要素と定義する．クラスタリング結果 $\mathcal{C}$ において，外れ要素と見なされたデータ要素の集合をノイズクラスタ $Noise_\mathcal{C}$ と定義する．クラスタリング結果 $\mathcal{C}$ からノイズクラスタを除いた集合 $\{C_1, \ldots, C_{N_\mathcal{C}}\}$ を $\mathcal{C}'$ とする．$\mathcal{C}'$ はクラスタの集合である．クラスタリング手法 CM は階層的クラスタリング手法と限定されるものではなく，入力パラメータの設定により様々なクラスタリング結果を出力可能な CHAMELEON [10]，OptiGrid [8]，O-Cluster [11]，そして FlexDice [15] などである．

2.2 対話的クラスタリングアルゴリズム

我々はユーザが自身の要求に近いクラスタを対話的に形成可能とすることを目的とした対話的クラスタリング手法 ICM を提案する．**図 1** に対話的クラスタリング手法 ICM の流れ，**図 2** に対話的クラスタリング手法 ICM の概要を示す．図 1 の斜線部分が対話的クラスタリング手法 ICM である．対話的クラスタリング手法 ICM はユーザ，クラスタリング手法 CM，特徴抽出手法 FEM，相違検出手法 DIFF，変換関数 TF，

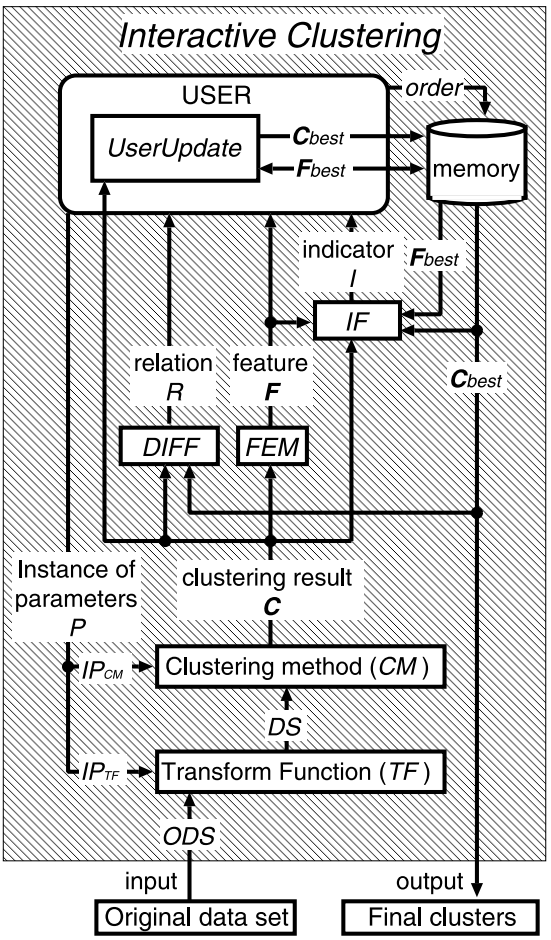

図 1 対話的クラスタリングの流れ
Fig. 1 Flow of the interactive clustering.

そして選択指標関数 IF によって構成され，入力はオリジナルデータセット ODS であり，出力はクラスタリング結果である．ここで相違検出手法 DIFF とは 2 つのクラスタリング結果 $\mathcal{C}_A$，$\mathcal{C}_B$ の関係 R を調べる手法である．相違検出手法 DIFF は以下の命題が真であるときの関係 R をユーザに出力する．

$$\forall C_A \in \mathcal{C}'_A, \forall C_B \in \mathcal{C}'_B$$
$$[(C_A, C_B) \in R \Leftrightarrow C_A \cap C_B \neq \emptyset]$$

変換関数 TF（Transform Function）とは，オリジナルデータセット ODS をクラスタリング手法 CM によってユーザの要求するクラスタリング結果を出力しやすいデータセット DS に変換する関数である．変換関数 TF はユーザの要求において重要な属性や属性間の相関関係をユーザが認識したときに，各属性の重要度や属性間の相関関係を考慮したメトリックを持つデータにオリジナルデータを変換することを可能にする関数である．変換前のデータセット ODS のデータ

Algorithm *ICM*(Original data Set: *ODS*);
begin
$\mathcal{C}_{best} =: \emptyset$; $\mathcal{F}_{best} =: \emptyset$;
while not($\mathcal{C}_{best}$_satisfies_user) **do**
begin
$IP_{CM}, IP_{TF} =: UserInputParameter()$;
$DS =: TF(ODS, IP_{TF})$;
$\mathcal{C}, \mathcal{C}' =: CM(DS, IP_{CM})$;
$\mathcal{F} =: FEM(\mathcal{C}')$;
$R =: DIFF(\mathcal{C}, \mathcal{C}_{best})$;
$I =: IF(\mathcal{C}, \mathcal{C}_{best}, \mathcal{F}, \mathcal{F}_{best})$;
$\mathcal{C}_{best} =: UserUpdateCluster(\mathcal{F}, \mathcal{F}_{best}, R, I)$;
$\mathcal{F}_{best} =: UserUpdateFeature(\mathcal{C}_{best})$;
end;
return($\mathcal{C}_{best}$);
end;

図 2 対話的クラスタリングアルゴリズム
Fig. 2 The interactive clustering.

要素と変換後のデータセット DS での対応するデータ要素をそれぞれ $1 \times dim$ 行列 $\boldsymbol{x}$，$\boldsymbol{y}$ とする．行列 $\boldsymbol{x}$，$\boldsymbol{y}$ はベクトルであり，その要素は実数とする．ユーザは変換関数 TF への入力パラメータである $dim \times dim$ 行列 $\boldsymbol{M}$ をパラメータ値の集合 IP_{TF} により設定し，変換関数 TF が出力する以下の式 (1) を満たす $\boldsymbol{y}$ にデータ要素 $\boldsymbol{x}$ を変換できる．

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x}^{\boldsymbol{T}} \quad (1)$$

以下で，図 2 に示す対話的クラスタリングアルゴリズムについて説明する．初期状態は最良クラスタリング結果 $\mathcal{C}_{best}$ と最良クラスタリング結果の特徴 $\mathcal{F}_{best}$ がともに空集合 $\emptyset$ である．$\mathcal{C}_{best}$ は対話的クラスタリングの実行過程においてユーザが最良と判断したクラスタリング結果であるものとし，$\mathcal{F}_{best}$ は $\mathcal{C}_{best}$ に対応するクラスタリング結果の特徴であるものとする．ユーザが $\mathcal{C}_{best}$ に満足するまでユーザとクラスタリング手法 CM が繰り返し情報交換しながら $\mathcal{C}_{best}$ を更新する．変換関数 TF にはオリジナルデータセット ODS と入力パラメータである行列 $\boldsymbol{M}$ の設定例 IP_{TF} が入力され，変換されたデータセット DS を出力する．クラスタリング手法 CM には変換関数 TF によって変換されたデータセット DS とユーザによって与えられたクラスタリング手法への入力パラメータの設定例 IP_{CM} が入力され，クラスタリング手法 CM はクラスタリング結果 $\mathcal{C}$ を出力する．クラスタリング結果 $\mathcal{C}$ は相違検出手法 DIFF，特徴抽出手法 FEM，選択指標関数 IF，そしてユーザに与えられる．特徴抽出手法 FEM はクラスタリング結果 $\mathcal{C}$ の特徴 $\mathcal{F}$ をユーザと選択指標関数 IF に提供し，相違検出手法 DIFF は $\mathcal{C}_{best}$ と $\mathcal{C}$ のクラスタ間のデータ要素交差関係 R をユーザに提供する．選択指標関数 IF は選択指標 I として $\mathcal{C}_{best}$ と $\mathcal{C}$ のどちらが自然なクラスタリング結果なのかの判定をユーザに知らせる．ユーザは $\mathcal{F}$，$\mathcal{F}_{best}$，R，そして I から $\mathcal{C}_{best}$ と $\mathcal{C}$ で要求に近いクラスタリング結果を選択し，$\mathcal{C}_{best}$ を更新する．更新された $\mathcal{C}_{best}$ に対応する特徴 $\mathcal{F}$ を $\mathcal{F}_{best}$ とする．更新が終了し，ユーザが $\mathcal{C}_{best}$ に満足したら，対話的クラスタリング手法 ICM は $\mathcal{C}_{best}$ を最終的なクラスタリング結果として出力する．ユーザが満足していなければ，対話的クラスタリング手法 ICM は上記のプロセスを繰り返し実行する．

2.3 特徴抽出手法（FEM）

本節において，クラスタリング結果から特異な分布のクラスタを算出し，クラスタリング結果の特徴を抽出する特徴抽出手法 FEM を提案する[16]．一般に，クラスタリングによって得られた結果は概要をとらえており扱いやすくなることが多い．しかし，高次元データの場合，このことは必ずしも成り立たない．本稿で提案する特徴抽出手法の目的は，特徴的な情報を自動的に抽出し，ユーザの高度な判断を支援することにある．特徴抽出手法への入力はクラスタリング手法 CM から得られたクラスタリング結果 $\mathcal{C}$ に対応するクラスタ集合 $\mathcal{C}'$ であり，特徴抽出手法 FEM の出力はクラスタリング結果 $\mathcal{C}$ の特徴 $\mathcal{F}$ である．クラスタリング結果 $\mathcal{C}$ の特徴 $\mathcal{F}$ には，各属性において特異な分布を持つクラスタが存在する場合，それらのクラスタとその属性におけるデータ要素の分布が記される．

特徴抽出手法 FEM では，特異な分布のクラスタを抽出するために外れ要素検出手法を用いる．外れ要素検出手法とはデータセットから類似した要素集合が形成され難い要素を検出する手法である．本稿では要素がベクトルとして表されるデータセットから外れ要素を検出するため，外れ要素検出手法を外れベクトル検出手法と呼ぶこととする．クラスタ集合 $\mathcal{C}' = \{C_1, \ldots, C_{N_\mathcal{C}}\}$ が特徴抽出手法 FEM に入力されたときの外れベクトル検出手法への入力は，各属性 d（$1 \leq d \leq dim$）に関する各クラスタ C_i の分布を表すベクトル集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leq i \leq N_\mathcal{C}\}$ である．$\boldsymbol{V}(C_i, d)$ をクラスタベクトルと呼ぶこととする．特徴抽出手法の入力データの属性 d の領域のサイズは正定数以下であるものと仮定する．値の範囲が連続である場合や範囲が広すぎる場合は離散化して自然数の値を割り当てる．各属性の領域内の値を値識別子

（Value ID）と呼ぶ．属性 d における値識別子の最大値を $m(d)$ としたとき，クラスタベクトル $\boldsymbol{V}(C_i, d)$ は以下の式 (2) で表される．

$$\boldsymbol{V}(C_i, d) = \frac{100}{N_{C_i}}(v(i,d,1), \ldots, v(i,d,m(d))) \quad (2)$$

ここで N_{C_i} はクラスタ C_i に含まれる全データ要素数であり，$v(i,d,j)$ はクラスタ C_i に含まれ，かつ，属性 d における値識別子が j であるデータ要素数である．

外れベクトル検出手法は，属性 d におけるクラスタベクトルの集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leq i \leq N_{\mathcal{C}}\}$ から他のクラスタベクトルと類似していないクラスタベクトルの集合を出力する．各属性 d（$1 \leq d \leq dim$）において，クラスタベクトルの集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leq i \leq N_{\mathcal{C}}\}$ を外れベクトル検出手法に入力することで，クラスタリング結果 $\mathcal{C}$ の特徴として特異なベクトル集合 $\mathcal{F}$ を特徴抽出手法は抽出する．外れベクトル検出手法には，一般的なデータ分析に対する要求と同様に容易な設定，少ない処理コスト，そして高い精度が要求される．外れベクトル検出手法として，上記の条件を満たしデータセットから外れ要素を検出可能な手法として FlexDice [15]，DBSCAN [5]，そして LOF [4] などがあげられる．

2.4 選択指標関数（IF）

クラスタリング手法 CM によって出力された 2 つのクラスタリング結果において，ユーザはクラスタリング結果やクラスタリング結果の特徴を見比べて 1 つのクラスタリング結果を選択するよりも，単純に一般的な観点から，より自然なクラスタリング結果と判断できる結果を選択したいことがあるだろう．ここで，自然なクラスタリングとは類似したデータ要素を同じクラスタに集め，類似していないデータ要素を別のクラスタに分類することである．本節では，クラスタリング結果やその特徴をユーザが詳細に解析することなく，2 つのクラスタリング結果から，より自然なクラスタリング結果を選択可能とすることを目的とした選択指標関数 IF を提案する．よって，選択指標関数 IF の導入により，一般的な観点から自然なクラスタリング結果をユーザに提示できれば，ユーザが理解し難い高次元なデータのクラスタリング結果の性質やユーザの要求とクラスタリング結果の性質の一致度合いを理解することなく，ユーザの要求するクラスタリング結果が得やすくなる．

選択指標関数 IF は 2 つのクラスタリング結果から 1 つを選択する指標となる選択指標 I を算出する関数である．選択指標関数 IF には 2 つのクラスタリング結果 $\mathcal{C}_A = \{C_{A1}, \ldots, C_{AN_{\mathcal{C}}}, Noise_{\mathcal{C}_A}\}$，$\mathcal{C}_B = \{C_{B1}, \ldots, C_{BN_{\mathcal{C}}}, Noise_{\mathcal{C}_B}\}$ に対応するノイズクラスタを除いたクラスタ集合 $\mathcal{C}'_A = \{C_{A1}, \ldots, C_{AN_{\mathcal{C}}}\}$，$\mathcal{C}'_B = \{C_{B1}, \ldots, C_{BN_{\mathcal{C}}}\}$ とそれらに対応する特徴抽出手法 FEM で得られた特徴 $\mathcal{F}_A$，$\mathcal{F}_B$ が入力される．選択指標関数 IF は $\mathcal{C}_A$ と $\mathcal{C}_B$ のどちらが自然なクラスタであるかの判定結果を以下の規則により選択指標 I を用いてユーザに示す．

- $I = 1$ のとき $\mathcal{C}_A$ は $\mathcal{C}_B$ より自然なクラスタリング結果である．
- $I = 2$ のとき $\mathcal{C}_B$ は $\mathcal{C}_A$ より自然なクラスタリング結果である．
- $I = 3$ のとき $\mathcal{C}_A$ と $\mathcal{C}_B$ のどちらが自然なクラスタリング結果か判定できない．

$I = 3$ の場合は選択指標 I によってクラスタリング結果を選択できないため，クラスタ数や外れ要素数や特徴の違いに基づいて，ユーザがどちらが良いのか判断することが要求される．

2.4.1 選択指標

本項では，与えられた 2 つのクラスタリング結果 $\mathcal{C}_A = \{C_{A1}, \ldots, C_{AN_{\mathcal{C}_A}}, Noise_{\mathcal{C}_A}\}$，$\mathcal{C}_B = \{C_{B1}, \ldots, C_{BN_{\mathcal{C}_B}}, Noise_{\mathcal{C}_B}\}$ に対し，選択指標 I の決定方法の概要を示し，その後で詳細な定義，方法について説明する．

我々の選択指標を算出するうえで重要な着想は，各クラスタに豊富な特徴を持つ結果が自然なクラスタリング結果であるという仮定に基づき，2 つのクラスタリング結果を比較することである．我々の着想に基づいた指標算出方法の概要を以下に示す．我々ははじめに 2 つのクラスタリング結果におけるクラスタの対応関係を調べ，$\mathcal{C}'_A$ の部分集合 $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{C}'_B$ の部分集合 $\mathcal{S}_l(\mathcal{C}'_B)$ を各クラスタ集合に属するクラスタに属するデータ要素の和がほぼ等しくなるように対応付ける．次に対応付けられたペアの数を L としたとき，L 個のクラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ において，クラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$，$\mathcal{S}_l(\mathcal{C}'_B)$ に関して，属するクラスタの特徴の和集合 $Att(\mathcal{S}_l(\mathcal{C}'_A))$, $Att(\mathcal{S}_l(\mathcal{C}'_B))$ の包含関係，交差関係を調べる（$1 \leq l \leq L$）．最後に，$Att(\mathcal{S}_l(\mathcal{C}'_A)) \supset Att(\mathcal{S}_l(\mathcal{C}'_B))$ が成り立つペアの数を α，$Att(\mathcal{S}_l(\mathcal{C}'_A)) \subset Att(\mathcal{S}_l(\mathcal{C}'_B))$ が成り立つペアの数を β，$Att(\mathcal{S}_l(\mathcal{C}'_A)) - Att(\mathcal{S}_l(\mathcal{C}'_B)) \neq \emptyset$ かつ $Att(\mathcal{S}_l(\mathcal{C}'_B)) - Att(\mathcal{S}_l(\mathcal{C}'_A)) \neq \emptyset$ が成り立つペアの数を γ，$Att(\mathcal{S}_l(\mathcal{C}'_A)) = Att(\mathcal{S}_l(\mathcal{C}'_B))$ が成り立つペアの数を δ として，α，β，γ，δ よりクラスタリング結果 $\mathcal{C}_A$，$\mathcal{C}_B$ のどちらが豊富な特徴を持っているかを判定

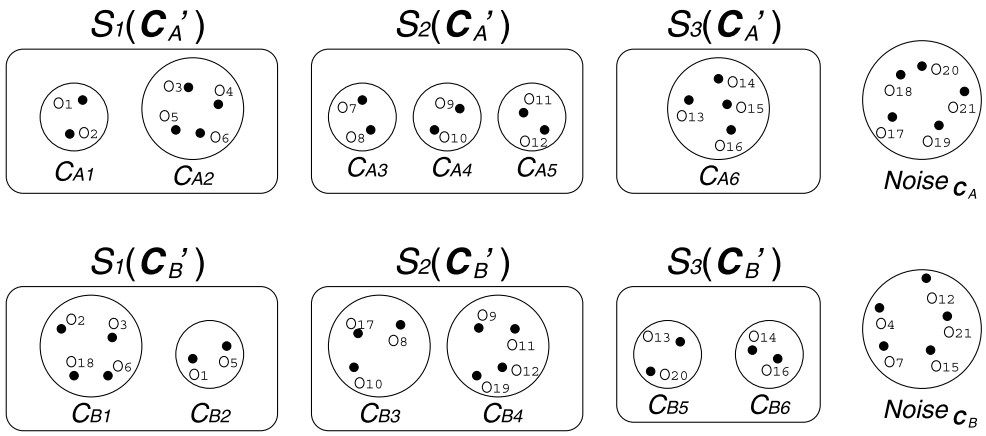

図 3 クラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{S}_l(\mathcal{C}'_B)$ の作成例
Fig. 3 An example of cluster set $\mathcal{S}_l(\mathcal{C}'_A)$ and $\mathcal{S}_l(\mathcal{C}'_B)$.

する．

1 つのクラスタリング結果における特徴が，もう 1 つのクラスタリング結果において特徴となるかどうか調べる．そのために，**図 3** のように，2 つのクラスタリング結果，つまり 2 つのクラスタ集合をそれぞれ分割し，クラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ とする．ただし，2 つのクラスタリング結果 $\mathcal{C}_A$ と $\mathcal{C}_B$ に関して，以下の式 (3) を満たすように各クラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{S}_l(\mathcal{C}'_B)$ を求める．

$$\bigcup_{C_X \in \mathcal{S}_l(\mathcal{C}'_A)} C_X - Noise_{\mathcal{C}_B} = \bigcup_{C_Y \in \mathcal{S}_l(\mathcal{C}'_B)} C_Y - Noise_{\mathcal{C}_A} \quad (3)$$

つまり，クラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$ に属するクラスタ C_X の和集合に対応するデータ要素群からノイズクラスタ $Noise_{\mathcal{C}_B}$ に属するデータ要素を除いたデータ要素群とクラスタ集合 $\mathcal{S}_l(\mathcal{C}'_B)$ に属するクラスタ C_Y の和集合に対応するデータ要素群からノイズクラスタ $Noise_{\mathcal{C}_A}$ に属するデータ要素を除いたデータ要素群が等しくなるようにクラスタ集合を対応付ける．以後では，上記の対応付けを持つクラスタ集合のペアを求める手法について説明する．

上述のクラスタの集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ を求める手続きを以下に示す $(1 \le l \le L)$．まず，与えられた 2 つのクラスタリング結果におけるノイズクラスタを除いたクラスタ集合 $\mathcal{C}'_A = \{C_{A1}, C_{A2}, \ldots, C_{AN_{\mathcal{C}_A}}\}$，$\mathcal{C}'_B = \{C_{B1}, C_{B2}, \ldots, C_{BN_{\mathcal{C}_B}}\}$ に対し，節点集合 $W_A \cup W_B$，枝の集合 E であるグラフ $G(W_A \cup W_B, E)$ を定義する．ここで，W_A は集合 $\{w_{A1}, w_{A2}, \ldots, w_{A_{\mathcal{C}_A}}\}$ とし，W_B は集合 $\{w_{B1}, w_{B2}, \ldots, w_{B_{\mathcal{C}_B}}\}$ とする．

クラスタ集合 $\mathcal{C}'_A$.$\mathcal{C}'_B$ とグラフ G の節点集合間の関係を定義する．全単射な写像 $Node$ を直積集合 $(\mathcal{C}'_A \cup \mathcal{C}'_B) \times (W_A \cup W_B)$ の部分集合 $\{(C_{Ai}, w_{Ai}) \mid 1 \le i \le N_{C_A}, C_{Ai} \in \mathcal{C}'_A \wedge w_{Ai} \in W_A\} \cup \{(C_{Bi}, w_{Bi}) \mid 1 \le i \le N_{C_B}, C_{Bi} \in \mathcal{C}'_B \wedge w_{Bi} \in W_B\}$ とする．また，$(C, w) \in Node$ を満たすとき，$w = Node(C)$，$C = Node^{-1}(w)$ で表すこととする．次に，クラスタ集合 $\mathcal{C}'_A$，$\mathcal{C}'_B$ とグラフ G の枝集合間の関係を定義する．枝集合 E は集合 $\{(Node(C_{Ai}), Node(C_{Bj})) \mid C_{Ai} \cap C_{Bj} \ne \emptyset, 1 \le i \le N_{\mathcal{C}_A}, 1 \le j \le N_{\mathcal{C}_B}\}$ とする．グラフ $G(W_A \cup W_B, E)$ は L 個の $W_A \cap W_l \ne \emptyset$ と $W_B \cap W_l \ne \emptyset$ を両方満たす連結成分 $G'_l(W_l, E_l)$ から構成されていると仮定する $(1 \le l \le L)$．

各 l $(1 \le l \le L)$ において，l 番目のグラフ G の連結成分 G'_l よりクラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$，$\mathcal{S}_l(\mathcal{C}'_B)$ を定義する．グラフ G'_l より，集合 $\mathcal{S}_l(\mathcal{C}'_A) = \{C \mid C = Node^{-1}(w), w \in W_l \cap W_A\}$，$\mathcal{S}_l(\mathcal{C}'_A) = \{C \mid C = Node^{-1}(w), w \in W_l \cap W_B\}$ である各クラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$，$\mathcal{S}_l(\mathcal{C}'_B)$ を求める．また，あるクラスタリング結果 $\mathcal{C}$ とあるクラスタリング集合 $\mathcal{S}$ が与えられたとき，特徴抽出手法 FEM によって，クラスタ集合 $\mathcal{S}$ に属するクラスタが特徴として検出された属性の集合を $Att(\mathcal{S})$ とする．上記により求まった L 個のクラスタ集合のペア $(\mathcal{S}_1(\mathcal{C}'_A), \mathcal{S}_1(\mathcal{C}'_B)), \ldots, (\mathcal{S}_L(\mathcal{C}'_A), \mathcal{S}_L(\mathcal{C}'_B))$ からなる集合をクラスタリング結果 $\mathcal{C}'_A$，$\mathcal{C}'_B$ の対応関係 CR と定義する．

上記で示した 2 つのクラスタリング結果 $\mathcal{C}'_A$，$\mathcal{C}'_B$ の対応関係 CR に基づく選択指標 I の決定方法を以下で説明する．選択指標 I の値は先に定義した α，β，γ，δ の値によって決定する．α，β，γ，δ の中で α

が最も大きいとき $I=1$，β が最も大きいとき $I=2$，γ，または，δ が最も大きいとき $I=3$ とする．

α の値は，対応関係 CR を定義している $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{S}_l(\mathcal{C}'_B)$ において，$\mathcal{S}_l(\mathcal{C}'_B)$ よりも $\mathcal{S}_l(\mathcal{C}'_A)$ の方が自然なクラスタ集合であると見なされたペアの数であると，我々は以下の推論により解釈する．$Att(\mathcal{S}_l(\mathcal{C}'_A)) \supset Att(\mathcal{S}_l(\mathcal{C}'_B))$ である場合，$\mathcal{S}_l(\mathcal{C}'_B)$ に属するクラスタの特徴は $\mathcal{S}_l(\mathcal{C}'_A)$ に属するクラスタの特徴としてすべて含まれており，かつ，$\mathcal{S}_l(\mathcal{C}'_A)$ に属するクラスタしか持たない特徴がある．そこで我々は以下の 2 つの仮定のどちらかが成り立っていると推測する．

- $\mathcal{S}_l(\mathcal{C}'_A)$ では類似した性質を持つデータ要素が同一のクラスタに属しているが，$\mathcal{S}_l(\mathcal{C}'_B)$ では別々のクラスタに分類されている．
- $\mathcal{S}_l(\mathcal{C}'_A)$ では類似していない性質を持つデータ要素が別々のクラスタに分類されているが，$\mathcal{S}_l(\mathcal{C}'_B)$ では同一のクラスタに属している．

β の値は，$\mathcal{S}_l(\mathcal{C}'_A)$ よりも $\mathcal{S}_l(\mathcal{C}'_B)$ の方が自然なクラスタ集合であると見なされたペアの数であると，我々は α と同様な推論により解釈する．

対応関係 CR に属するクラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ において，$Att(\mathcal{S}_l(\mathcal{C}'_A)) - Att(\mathcal{S}_l(\mathcal{C}'_B)) \neq \emptyset$，かつ，$Att(\mathcal{S}_l(\mathcal{C}'_B)) - Att(\mathcal{S}_l(\mathcal{C}'_A)) \neq \emptyset$ の場合，$Att(\mathcal{S}_l(\mathcal{C}'_A)) \supset Att(\mathcal{S}_l(\mathcal{C}'_B))$ において仮定する 2 つの状況，または，$Att(\mathcal{S}_l(\mathcal{C}'_A)) \subset Att(\mathcal{S}_l(\mathcal{C}'_B))$ において仮定する 2 つの状況のうちのどちらかの状況であると推測する．γ の値はクラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ において，どちらが自然なクラスタを形成できているのか分からないペアの数と解釈する．

クラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ において，$Att(\mathcal{S}_l(\mathcal{C}'_A)) = Att(\mathcal{S}_l(\mathcal{C}'_B))$ である場合，$\mathcal{S}_l(\mathcal{C}'_A)$ に属するクラスタと $\mathcal{S}_l(\mathcal{C}'_B)$ に属するクラスタは同一の性質を持つと推測する．したがって，δ の値は $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{S}_l(\mathcal{C}'_B)$ のどちらが自然なクラスタを形成できているのか分からず，また，どちらのクラスタリング結果も類似した結果であるペアの数であると解釈する．

2.4.2 選択指標関数の動作例

本項では図 3 のように，2 つのクラスタリング結果，クラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ が作成されたとき，選択指標関数の動作例を示す．

図 3 において，クラスタ C_{A1} は属性 1，2 で，C_{A2} は属性 2 で，C_{A3} は属性 2 で，C_{A5} は属性 3 で，C_{A6} は属性 2，3 で特徴が現れたとする．また，クラスタ C_{B1} は属性 2 で，C_{B3} は属性 1 で，C_{B4} は属性 3 で，C_{B6} は属性 3 で特徴が現れたとする．このとき，$Att(\mathcal{S}_1(\mathcal{C}'_A)) = \{1,2\}$，$Att(\mathcal{S}_2(\mathcal{C}'_A)) = \{2,3\}$，$Att(\mathcal{S}_3(\mathcal{C}'_A)) = \{2,3\}$，$Att(\mathcal{S}_1(\mathcal{C}'_B)) = \{2\}$，$Att(\mathcal{S}_2(\mathcal{C}'_B)) = \{1,3\}$，$Att(\mathcal{S}_3(\mathcal{C}'_B)) = \{3\}$ である．

上記の例のとき，各 $Att(\mathcal{S}_l(\mathcal{C}'_A))$ と $Att(\mathcal{S}_l(\mathcal{C}'_B))$ の関係は $Att(\mathcal{S}_1(\mathcal{C}'_A)) \supset Att(\mathcal{S}_1(\mathcal{C}'_B))$，$Att(\mathcal{S}_2(\mathcal{C}'_A)) - Att(\mathcal{S}_2(\mathcal{C}'_B)) \neq \emptyset$ かつ $Att(\mathcal{S}_2(\mathcal{C}'_B)) - Att(\mathcal{S}_2(\mathcal{C}'_A)) \neq \emptyset$，$Att(\mathcal{S}_3(\mathcal{C}'_A)) \supset Att(\mathcal{S}_3(\mathcal{C}'_B))$ となる．したがって，$\alpha=2$，$\beta=0$，$\gamma=1$，$\delta=0$ となり，α，β，γ，δ の中で最も α が大きいので，$I=1$ となる．選択指標関数は結果 $\mathcal{C}_B$ よりも結果 $\mathcal{C}_A$ が自然な結果であるという情報をユーザに提示する．

3. 意思決定支援の具体例

本章では**図 4** に示す合成 2 次元データを用いて意思決定支援の具体例を示す．図 4 は対話的クラスタリング手法 ICM へ入力するオリジナルデータセット *ODS* を変換関数 TF によって変換したデータセット *DS* の描画である．データセット *DS* は 1 つの点が 1 つのデータ要素，横軸が属性 1（属性 x と呼ぶ），縦軸が属性 2（属性 y と呼ぶ）を表している．図 4 のように低次元データでは形成されるクラスタの形状を容易に把握しやすいが，高次元データではクラスタの形状を把握することは難しい．本章では形成されたクラスタを視覚的に把握しやすくするために，2 次元の描画データセット *DS* を使用する．3.1 節ではクラスタリング手法の入力パラメータについて説明する．3.2 節では図 4 の入力に対する特徴抽出例，3.3 節では意思決定支援の具体例を示す．

3.1 クラスタリング手法の入力パラメータ

本節では 2 章で提案した対話的クラスタリング手法におけるクラスタリング手法，および，特徴抽出手法で一例として使用可能なクラスタリング手法 FlexDice [15] の入力パラメータについて説明する．クラスタリング手法 FlexDice はトップダウンにデータ空間を不均一

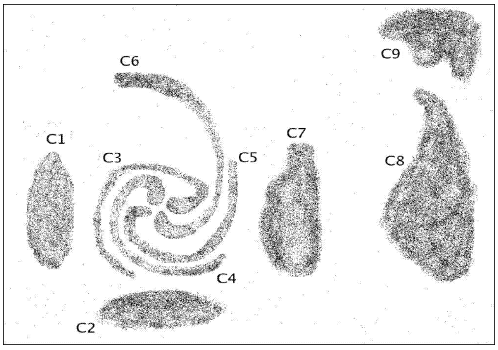

図 4 2 次元入力データ（データ要素数：59,028）
Fig. 4 2-dimensional input data set.

なサイズのセルに分割することを基本とするクラスタリング手法である．ただし，提案する対話的クラスタリング手法におけるクラスタリング手法や提案する特徴抽出手法では必ずしもクラスタリング手法 FlexDice を使用する必要はない．

クラスタリング手法 FlexDice は 4 つの入力パラメータを持つが，本稿ではクラスタリング結果を変化させるための入力パラメータとして P_{min}，P_b，そして P_{ele} を使用する．P_{min} は要素数によってセル内のデータ要素を外れ要素とするのかしないのかを決めるパラメータである．P_{min} の値を大きくすると，P_{min} の値が小さいときに形成されるクラスタが分割され，小さなクラスタが形成されやすくなる．また，P_{min} の値を大きくすると，一般に外れ要素数が増えるパラメータである．P_b はセルの分割回数の最大値を決めるパラメータである．P_b は P_{min} と同様であり，P_b の値を大きくすると，P_b の値が小さいときに形成されるクラスタが分割され，小さなクラスタが形成されやすくなる．また，P_b は値を大きくすると外れ要素数が増えるパラメータである．P_{ele} はクラスタに含まれる要素数に応じてそのクラスタをクラスタ，または，ノイズクラスタであるかを決めるパラメータである．P_{ele} は値を大きくするとクラスタ数が減り，かつ，外れ要素数が増えるパラメータである．

クラスタリング手法 CM が OptiGrid [8] であるとき，OptiGrid に入力するパラメータ値の集合にはデータ空間の分割面を定める値や 1 つの部分空間を分割するときの分割面の数を含む．

3.2 特徴抽出例

特徴抽出手法 FEM がクラスタ集合から各属性に対してデータ要素の分布が特異なクラスタ群を抽出する様子を例示する．本節では，特徴抽出手法 FEM への入力を図 4 の視覚的に明らかな 9 個のクラスタの集合 $\{Ci \mid 1 \leq i \leq 9\}$ とする．実際に図 4 のデータセット DS とある入力パラメータ値をクラスタリング手法 FlexDice に入力するとクラスタ集合 $\{Ci \mid 1 \leq i \leq 9\}$ を出力することができた．以下では，特徴抽出手法 FEM によって，$C1$ と $C7$ が属性 x に特徴を持ち，$C2$，$C6$，$C9$ が属性 y に特徴を持つことを自動で抽出できることを示す．

はじめに各クラスタ Ci における各属性 d に関して，各値識別子に対応する値域に値を持つデータ要素数を求める．ここでは均等に属性 x と属性 y の値域を 10 等分し，属性 x，y の値識別子をそれぞれ 1 から 10 までの整数とする．次に，属性 x，y において，各クラスタ Ci に含まれるデータ要素数に対する各値識別子に対応するデータ要素数の割合をベクトルとして算出する．算出されたベクトルは**図 5**，**図 6** のように図示できる．図 5 と図 6 の横軸は値識別子であり，縦軸はクラスタ Ci に含まれるデータ要素に対する各値識別子に属するデータ要素数の百分率である．

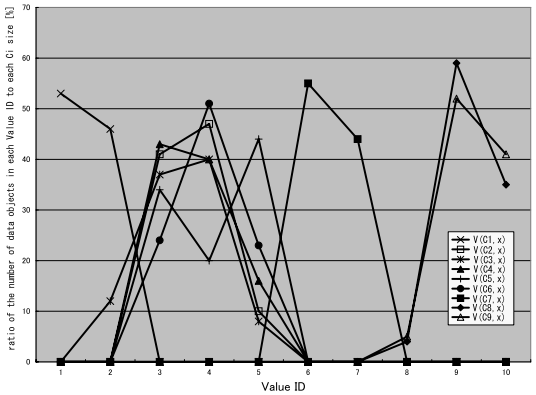

図 5 C3，C4，C5，そして C6 が別々のクラスタとなったときの属性 x に関する各クラスタのクラスタベクトル
Fig. 5 Vectors in each cluster for attribute x, when C3, C4, C5 and C6 are not one cluster.

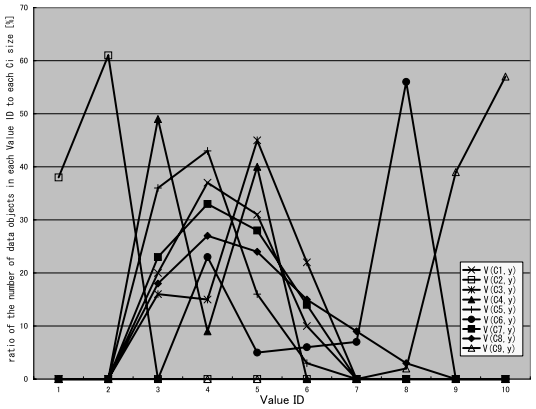

図 6 C3，C4，C5，そして C6 が別々のクラスタとなったときの属性 y に関する各クラスタのクラスタベクトル
Fig. 6 Vectors in each cluster for attribute y, when C3, C4, C5 and C6 are not one cluster.

図 5，図 6 において，折れ線により表現されている各クラスタベクトル $\boldsymbol{V}(Ci, d)$ は，図 4 の各クラスタ Ci に対応し，外れベクトル検出手法の入力データ要素である．外れベクトル検出手法への入力データの属性はクラスタベクトル $\boldsymbol{V}(Ci, d)$ の値識別子である．外れベクトル検出手法への入力は 2 つのデータセット $\{\boldsymbol{V}(C1, x), \boldsymbol{V}(C2, x), \ldots, \boldsymbol{V}(C9, x)\}$，$\{\boldsymbol{V}(C1, y), \boldsymbol{V}(C2, y), \ldots, \boldsymbol{V}(C9, y)\}$ である．各データセットから外れベクトル検出手法を用いて外れベクトルを検出する．外れベクトル検出手法にクラスタリン

グ手法 FlexDice を使用したとき，図 5 の属性 x では，$\boldsymbol{V}(C2,x)$，$\boldsymbol{V}(C3,x)$，$\boldsymbol{V}(C4,x)$，そして $\boldsymbol{V}(C6,x)$ を含むクラスタと，$\boldsymbol{V}(C8,x)$ と $\boldsymbol{V}(C9,x)$ を含むクラスタを形成した．$\boldsymbol{V}(C1,x)$ と $\boldsymbol{V}(C7,x)$ は外れベクトルとして分類された．したがって，属性 x に関して外れベクトル検出手法は $\boldsymbol{V}(C1,x)$ と $\boldsymbol{V}(C7,x)$ を外れベクトルとして出力する．同様に図 6 の属性 y に関してクラスタリング手法 FlexDice を用いて外れベクトルを検出したとき，外れベクトル検出手法は $\boldsymbol{V}(C2,y)$，$\boldsymbol{V}(C6,y)$，そして $\boldsymbol{V}(C9,y)$ を外れベクトルとして出力する．他のクラスタベクトル $\boldsymbol{V}(C1,y)$，$\boldsymbol{V}(C3,y)$，$\boldsymbol{V}(C4,y)$，$\boldsymbol{V}(C5,y)$，$\boldsymbol{V}(C7,y)$，そして $\boldsymbol{V}(C8,y)$ は 1 つのクラスタを形成した．

以上の結果から，属性 x に関して図 5 に示すような特異な値の分布を持つクラスタ $C1$ と $C7$，属性 y に関して図 6 に示すような特異な値の分布を持つクラスタ $C2$，$C6$，そして $C9$ が図 4 をクラスタリングしたときの特徴として特徴抽出手法は出力する．

3.3 特徴抽出手法を用いた意思決定支援例

3.2 節では $C3$，$C4$，$C5$，そして $C6$ の 4 つのクラスタが識別された例を使用したが，クラスタリング手法 CM への入力パラメータを変化させたとき，クラスタリング手法 CM は距離が近いこれらのクラスタを 1 つに統合する．実際に図 4 のデータセット DS とある入力パラメータをクラスタリング手法 FlexDice に入力した場合，$C3$，$C4$，$C5$，そして $C6$ の 4 つのクラスタに含まれるデータ要素を 1 つのクラスタとし，他の $C1$，$C2$，$C7$，$C8$，そして $C9$ をそれぞれ 1 つのクラスタとして出力した．入力パラメータ設定例 IP_A をクラスタリング手法に入力した場合，$C3$，$C4$，$C5$，そして $C6$ の 4 つクラスタに含まれるデータ要素が 1 つのクラスタ $C10$ として形成されたときのクラスタリング結果を $\mathcal{C}_A$ とする．また，3.2 節のように $C3$，$C4$，$C5$，そして $C6$ が 4 つのクラスタとして形成された場合，クラスタリング結果を $\mathcal{C}_B$，またそのときの入力パラメータ設定例を IP_B とする．結果 $\mathcal{C}_A$ において，属性 x と属性 y に関する各クラスタに含まれるデータ要素の分布をそれぞれ**図 7**，**図 8** に示す．図 7，図 8 の右下に凡例を示す．クラスタリング手法 FlexDice を外れベクトル検出手法として用いると，属性 x において $\boldsymbol{V}(C1,x)$ と $\boldsymbol{V}(C7,x)$ が外れベクトルとして検出され，属性 y において $\boldsymbol{V}(C2,y)$ と $\boldsymbol{V}(C9,y)$ が外れベクトルとして検出された．ここで，クラスタ集合 $\mathcal{S}$ はクラスタ C と以下のように対応付けられたとする．

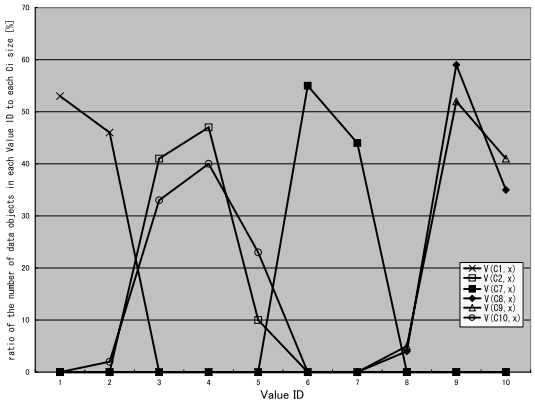

図 7 C3，C4，C5，そして C6 が 1 つのクラスタ C10 となったときの属性 x に関する各クラスタのクラスタベクトル
Fig. 7 Vectors in each cluster for attribute x, when C3, C4, C5 and C6 are one cluster C10.

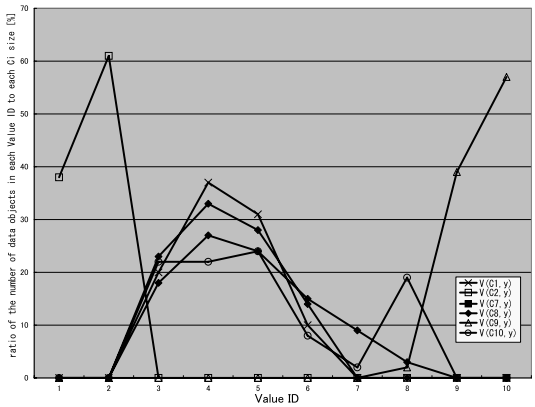

図 8 C3，C4，C5，そして C6 が 1 つのクラスタ C10 となったときの属性 y に関する各クラスタのクラスタベクトル
Fig. 8 Vectors in each cluster for attribute y, when C3, C4, C5 and C6 are one cluster C10.

- $\mathcal{S}_1(\mathcal{C}'_A) = \{C1\}$，$\mathcal{S}_2(\mathcal{C}'_A) = \{C2\}$，$\mathcal{S}_3(\mathcal{C}'_A) = \{C10\}$，$\mathcal{S}_4(\mathcal{C}'_A) = \{C7\}$，$\mathcal{S}_5(\mathcal{C}'_A) = \{C8\}$，$\mathcal{S}_6(\mathcal{C}'_A) = \{C9\}$
- $\mathcal{S}_1(\mathcal{C}'_B) = \{C1\}$，$\mathcal{S}_2(\mathcal{C}'_B) = \{C2\}$，$\mathcal{S}_3(\mathcal{C}'_B) = \{C3, C4, C5, C6\}$，$\mathcal{S}_4(\mathcal{C}'_B) = \{C7\}$，$\mathcal{S}_5(\mathcal{C}'_B) = \{C8\}$，$\mathcal{S}_6(\mathcal{C}'_B) = \{C9\}$

特徴抽出手法 FEM は特徴 $\mathcal{F}_A$，$\mathcal{F}_B$，相違検出手法 DIFF は $\mathcal{C}_A$ と $\mathcal{C}_B$ の関係 R，選択指標関数 IF は選択指標 I をユーザに提供する．以下に $\mathcal{F}_A$，$\mathcal{F}_B$，R，そして I の内容を示す．

$\mathcal{F}_A$： クラスタ $C1$ と $C7$ が属性 x で特異な分布を持ち，クラスタ $C2$ と $C9$ が属性 y で特異な分布を持つ．またそれらの分布は図 7，図 8 に示すとおりである．

$\mathcal{F}_B$： クラスタ $C1$ と $C7$ が属性 x で特異な分布を持ち，クラスタ $C2$，$C6$，そして $C9$ が属性 y で特異な分布を持つ．またそれらの分布は図 5，図 6 に示すとおりである．

R： クラスタリング結果 $\mathcal{C}_A$ における $C10$ はクラスタリング結果 $\mathcal{C}_B$ では 4 つのクラスタ $C3$，$C4$，$C5$，そして $C6$ として形成される．他の $C1$，$C2$，$C7$，$C8$，そして $C9$ は $\mathcal{C}_A$ と $\mathcal{C}_B$ で対応している．

I： $\mathcal{S}_i(\mathcal{C}'_A)$ と $\mathcal{S}_i(\mathcal{C}'_B)$ のすべてのペアから α，β，γ，δ を求めると，$\alpha = 0$，$\beta = 1$，$\gamma = 0$，$\delta = 5$ となり，δ が最も大きい（$1 \leq i \leq 6$）．したがって，$I = 3$，つまり選択指標関数 IF によって $\mathcal{C}_B$ は $\mathcal{C}_A$ よりも自然なクラスタを得ているが，違いがほとんどないと判断されたことを示す．

選択指標関数 IF が $I = 3$ を出力し，ユーザは選択指標関数 IF 以外から得られる情報を基に結果 $\mathcal{C}_A$，または，$\mathcal{C}_B$ を選択しなければならない．この場合は，選択指標関数 IF の出力である選択指標 I によって，$\mathcal{C}_A$，または，$\mathcal{C}_B$ のどちらかを選択することはできなかった．$\mathcal{C}_A$，または，$\mathcal{C}_B$ のどちらを選ぶかはユーザによって異なるだろう．また，選択した結果 $\mathcal{C}_A$，または，$\mathcal{C}_B$ で満足するかどうかもユーザによって異なるだろう．$\mathcal{C}_A$ では，すべての属性で特異な分布を持たなかった $C10$ が $\mathcal{C}_B$ では 4 つのクラスタに分割され，さらに分割されて形成された 1 つのクラスタ $C6$ が属性 y で特異な分布を持つことが分かる．図 6 から $C6$ は属性 y の値識別子 8 であるデータ要素を高い割合で含み，他のクラスタには属性 y の値識別子 8 であるデータ要素を高い割合で含むものが存在しないことも分かる．特異な分布を持つクラスタをできるだけ多く抽出したい応用の場合，ユーザは以上の情報から要求に近い結果は $\mathcal{C}_B$ であると判断するだろう．クラスタリング結果 $\mathcal{C}_B$ に満足した場合，対話的クラスタリング手法 ICM はクラスタリング結果 $\mathcal{C}_B$ を $\mathcal{C}_{best}$ として出力する．$\mathcal{C}_B$ に満足できない場合，例として以下の 2 つのパラメータ設定例が考えられる．

- 入力パラメータ設定例 IP_A よりも IP_B に近く IP_B とは異なる入力パラメータ設定例をユーザはクラスタリング手法 CM に入力する．
- クラスタリング結果 $\mathcal{C}_B$ に関して，特徴抽出手法は属性 x で 2 つのクラスタ，属性 y で 3 つのクラスタを特徴として検出した．特徴が多く検出される属性を重要な属性として次回のクラスタリングを実行するために，属性 y の尺度を大きく変換するようなパラメータ設定値 IP_{TF} を TF に入力する．

以上の例では，対話的クラスタリング手法 ICM により，ユーザはクラスタリング結果 $\mathcal{C}_B$ を選択しやすくなり，かつ，要求するクラスタ集合がまったく分からないときにでも 2 つのクラスタリング結果の特徴を比較することによりどちらの結果が注目に値する特徴を持っているかを判断するために用いる情報を得ることが可能になる．

4. 対話的クラスタリング手法 ICM のベンチマークデータへの適用

対話的クラスタリング手法 ICM の有効性をシミュレーション実験により示す．4.1 節において対話的クラスタリング手法 ICM の評価に使用する評価関数や入力データを説明する．4.2 節では，2 つのクラスタリング結果から 1 つの自然なクラスタリング結果をユーザが求める場合，選択指標関数 IF がどのくらい有効であるかを調べる．4.3 節では，ユーザが対話的にクラスタリング結果を得る過程と最終的に得た結果がユーザの要求に近い結果が得られるかどうかを調べる．本章のシミュレーション実験において，クラスタリング手法 CM として，FlexDice を使用した．本稿の実験において，データはクラシフィケーションにおけるクラスラベルを持つものとし，問題の簡単化のため，クラスラベルが “0”，または，“1” の 2 値であるデータセットのみを入力とした．

4.1 実験の準備

選択指標 I に基づいてクラスタリング結果を選択したとき，要求に近い結果を得られるかどうかを調べるために，クラスタリング結果とユーザの要求するクラスタリング結果の近さを算出する関数を定義する．クラスラベルがないデータセットにおいて，ユーザは同一のクラスラベルを持つであろうデータ要素を集めたいとする．データセットにおいて同一のクラスラベルを持つデータ要素が集められたかどうかを評価するために，分類エラー E_c がよく使用される[6]．E_C は入力データ要素全体に対して誤って分類された要素数の割合であり，0 に近いほど高精度な出力と評価できる．

クラスタリングアルゴリズムが要素数 $s_1, \ldots, s_{N_C}$ のクラスタを出力し，それぞれのクラスタの最多共通クラスのラベルを $L_1, \ldots, L_{N_C}$ とする．また，$s_1, \ldots, s_{N_C}$ の各クラスタの最多共通ラベルを持つ要素数を $m_1, \ldots, m_{N_C}$ としたとき，E_C は以下の式 (4)

によって定義される．

$$E_C = \frac{\sum_{i=1}^{N_C}(s_i - m_i)}{\sum_{i=1}^{N_C} s_i} = \frac{\sum_{i=1}^{N_C}(s_i - m_i)}{N} \quad (4)$$

我々は同じラベルを持つ多くのデータ要素を 1 つのクラスタに集められたことを高く評価したいだけでなく，割合が少ないラベルを持つデータ要素を多く集められたことも高く評価したい．しかし，分類エラー E_C ではクラスラベル間のデータ要素数の差が大きいデータセットの場合，割合が少ないラベルを持つデータ要素を集められたことについて高く評価されにくい．我々はバランスのとれていないデータセットであってもバランスのとれたデータセットとして評価可能なクラスタリング向け分類エラー $E'_{\mathcal{C}}$ を定義する．クラスタリング向け分類エラー $E'_{\mathcal{C}}$ は最良値が 0 であり，最悪値が 0.5 である．クラスタリング向け分類エラー $E'_{\mathcal{C}}$ の主クラスラベルは最多共通ラベルではなく，入力データのクラスラベルの各値を持つデータ要素が同数であったと仮定したときの比率が高いクラスラベルとする．

クラスタリング向け分類エラー $E'_{\mathcal{C}}$ をクラスラベルが “0” と “1” のときについて定義する．なお，クラスラベルの値の数は複数であっても容易に拡張できる．あるデータ要素集合 D が与えられたときに，D に属するクラスラベル “0” のデータ要素数を $N0(D)$ で，クラスラベル “1” のデータ要素数を $N1(D)$ で表すとする．このとき，入力データセット DS において，クラスラベルが “0” の全データ要素数は $N0(DS)$，クラスラベルが “1” の全データ要素数は $N1(DS)$ とする．クラスタリングアルゴリズムがクラスタリング結果 $\mathcal{C}$ を出力したとする．結果 $\mathcal{C}$ に属する任意のクラスタ C のデータ要素数に対する主クラスラベル以外のラベルを持つデータ要素数を $MCL(C)$ とするとき，クラスタリング向け分類エラー $E'_{\mathcal{C}}$ は以下の式 (5) によって定義される．ここで，N は入力データ DS のデータ要素数である．

$$E'_{\mathcal{C}} = \sum_{C \in \mathcal{C}} \frac{MCL(C)}{N0(C) \times \frac{N}{N0(DS)} + N1(C) \times \frac{N}{N1(DS)}} \quad (5)$$

本稿では，2 つのクラスタリング結果が与えられたとき，選択指標関数 IF によって選択指標 I 求め，選択指標 I に従って選択したクラスタリング結果が自然なクラスタリング結果であるかどうかをクラスタリング向け分類エラー $E'_{\mathcal{C}}$ を用いて評価する．

実験において使用するベンチマークデータは KDD アーカイブ[14] の “Census Income Database”（収入データ）である．収入データは 15 属性を含む 32,561 個のデータ要素から構成される．このデータセットは分類に関するデータであり 7,841 個のデータ要素と 24,720 個のデータ要素が 2 つのクラスラベルにより分類されている．クラスラベルはユーザや対話的クラスタリング手法 ICM には未知であるとして実験し，分類エラーを求めるためにのみ使用される．

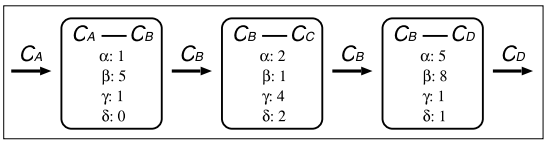

図 9 2 つの結果を比較して 1 つの結果を選択することを繰り返す過程
Fig. 9 The process of acquiring a desired clustering result.

4.2 選択指標関数 IF の妥当性

収入データにおいて，ランダムに入力パラメータ設定例 IP_{CM} を入力した 2 つのクラスタリング結果を比較する．比較した回数は 10 回である．選択指標関数 IF が $I = 3$ を出力し，自然なクラスタを選択する指標を示せなかった結果は 4 回であった．したがって，選択指標関数 IF は 60%の確率で二者択一を行った情報をユーザに与えることが可能であった．

選択指標関数 IF の選択どおりに 2 つのクラスタリング結果から 1 つのクラスタリング結果を選んだ場合，選択指標関数 IF によって E'_C を約 83%の確率で減少させることが可能であった．

4.3 対話的クラスタリング手法 ICM によるユーザへの支援の例

本節では収入データを用いて，ユーザが対話的にクラスタリング結果を得る過程を示す．2 つの結果を比較し 1 つの結果を選択することを繰り返した過程を**図 9** に示す．比較したクラスタリング結果を実線で結び，比較した結果の左に記された結果が 2.4 節の $\mathcal{C}_A$ に対応し，右に記された結果が $\mathcal{C}_B$ に対応する．2 つの結果を比較した後に出力される矢印の上に記した結果が選択した結果である．また，そのときに入力した入力パラメータ設定例と $E'_{\mathcal{C}}$ の値を**表 1** に示す．

(1) 2 つの結果 $\mathcal{C}_A$ と $\mathcal{C}_B$ を比較する．選択指標関数 IF を用いて α，β，γ，δ の値を算出すると図 9 に示す結果となった．β が最も大きい数であるので，IF は選択指標 $I = 2$（$\mathcal{C}_B$ が $\mathcal{C}_A$ より自然な結果）を出力する．ユーザは IF の選択指標に基づいて $\mathcal{C}_B$ を選択する．

(2) 前回の比較で選択した結果 $\mathcal{C}_B$ と新たにパラ

表 1 各結果における入力パラメータ値，E_C，E'_C の値
Table 1 The values of parameters, classification error E_C and clustering error E'_C on each clustering result.

	P_{min}	P_b	P_{ele}	clusters	E_C	E'_C
$\mathcal{C}_A$	1	2	5	24	0.225	0.356
$\mathcal{C}_B$	20	2	10	13	0.235	0.357
$\mathcal{C}_C$	10	2	20	12	0.233	0.342
$\mathcal{C}_D$	3	2	5	27	0.229	0.316

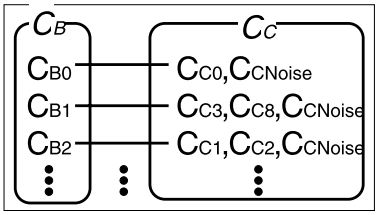

図 10 $\mathcal{C}_B$ と $\mathcal{C}_C$ におけるクラスタの関係
Fig. 10 Relation between clusters in $\mathcal{C}_B$ and ones in $\mathcal{C}_C$.

メータ P_{CM} を入力して得た結果 $\mathcal{C}_C$ を比較する．選択指標関数 IF を用いて α，β，γ，δ の値を算出すると図 9 に示す結果となった．γ が最も大きい数であるので，IF は選択指標 $I = 3$（$\mathcal{C}_B$ と $\mathcal{C}_C$ のどちらが自然な結果か判定できない）を出力する．ユーザは I 以外の情報を基に $\mathcal{C}_B$ か $\mathcal{C}_C$ を選択する．ユーザは DIFF から**図 10** に示す R と FEM から**図 11**，**図 12** に示す特徴 $\mathcal{F}_B$，$\mathcal{F}_C$ を得る．紙面の都合上 R，$\mathcal{F}_B$，そして $\mathcal{F}_C$ は 1 部のみを図に示した．図 10 には結果 $\mathcal{C}_B$ と $\mathcal{C}_C$ におけるクラスタ間の関係を示した．図 11 には結果 $\mathcal{C}_B$ において，属性 3 に特徴を持つクラスタとそのクラスタベクトルを示した．1 本の折れ線は 1 つのクラスタベクトルを表し，各クラスタに含まれるデータ要素が属性 3 において持つ値の分布を表す．図 12 は結果 $\mathcal{C}_C$ において，属性 3 に特徴を持つクラスタとそのクラスタベクトルを示した図である．

ユーザは図 11 と図 12 を見て属性 3 の値 5 が 30%以上ある C_{A2} ようなクラスタが欲しいと考え，結果 $\mathcal{C}_B$ を選択する．

(3) 前回の比較で選択した結果 $\mathcal{C}_B$ と新たにパラメータ設定例 IP_{CM} を入力して得た結果 $\mathcal{C}_D$ を比較する．新たに入力したパラメータ設定例 IP_{CM} は結果 $\mathcal{C}_B$ のときの入力よりもパラメータ P_{min} を小さくし外れ要素数を減らし，クラスタ数を増やす設定にした．また，P_{ele} を小さくすることで外れ要素数を減らす設定にした．IF を用いて α，β，γ，δ の値を算出すると図 9 に示す結果となった．β が最も大きい数であるので，IF は選択指標 $I = 2$（$\mathcal{C}_D$ が $\mathcal{C}_B$ より自然な結果）を出力する．ユーザは IF の選択指標に基づいて結果 $\mathcal{C}_D$ を選択する．

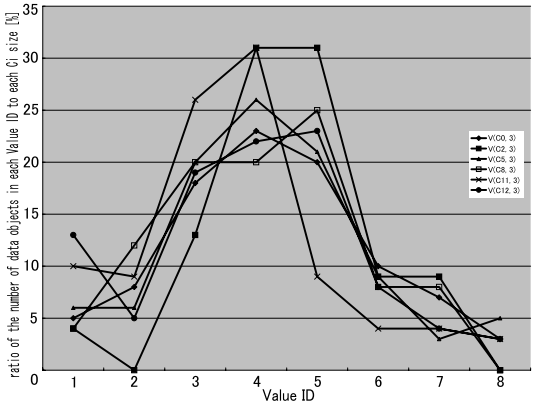

図 11 収入データの結果 $\mathcal{C}_A$ における属性 3 に関する特徴として検出されたクラスタベクトル（%）
Fig. 11 Characteristic vectors found for attribute 3 in clustering results $\mathcal{C}_A$ of Census Income Database.

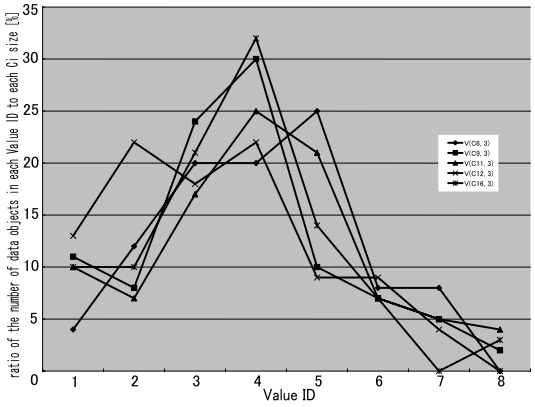

図 12 収入データの結果 $\mathcal{C}_B$ における属性 3 に関する特徴として検出されたクラスタベクトル（%）
Fig. 12 Characteristic vectors found for attribute 3 in clustering results $\mathcal{C}_B$ of Census Income Database.

以上のような 2 つの結果の比較と選択を繰り返し，要求に近いクラスタが抽出されたと判断したところでユーザは対話的クラスタリングを停止する．今回の例では最初に選択されていた結果 $\mathcal{C}_A$ から最後に選択した結果 $\mathcal{C}_D$ までで，E'_C を 0.356 から 0.316 まで減少させることができた．一方，式 (4) で示した従来の分類エラー E_C が大きく変化しなかった理由として，使用した収入データがアンバランスなデータだったためと考えられる．ユーザの要求する結果がラベルによって分かれた結果とすると，対話的クラスタリング手法

によって有効な結果を出力できた．

5. おわりに

本稿ではユーザの要求に近いクラスタを対話的に導くことを目標とした対話的クラスタリング手法とその構成要素である特徴抽出手法および選択指標関数を提案した．シミュレーション実験ではベンチマークデータを用いて対話的にクラスタリング結果を選択する過程と最終的に得られたクラスタリング結果が分類エラーの低減というユーザの要求に応じたクラスタリング結果であったことを示した．

さらに，系統的なパラメータ設定支援を行うため，どのパラメータをどのように変化させることで，どのようなクラスタリング結果が得られるかをユーザに提示し，よりユーザに優しいクラスタリング手法を開発することを今後の課題とする．また，ユーザが自身の要求を明確にしながら対話的に要求に近いクラスタを形成することを今後の目標とする．

参考文献

1) Aggarwal, C.C.: Towards meaningful high-dimensional nearest neighbor search by human-computer interaction, *Proc. 18th Int. Conf. on Data Engineering* (*ICDE '02*), pp.593–604 (2002).
2) Aggarwal, C.C.: Towards effective and interpretable data mining by visual interaction, *ACM-SIGKDD Explorations*, Vol.3, pp.11–22 (2002).
3) Aggarwal, C.C.: A human-computer interactive method for projected clustering, *IEEE Trans. Knowledge and Data Engineering*, Vol.16, No.4, pp.448–460 (2004).
4) Breunig, M.M., Kriegel, H.-P., Hg, R.T. and Sander, J.: LOF: Identifying density-based local outliers, *Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data* (*SIGMOD '00*), pp.93–104 (2000).
5) Ester, M., Kriegel, H.-P., Sander, J. and Xu, X.: A density-based algorithm for discovering clusters in large spatial Databases with Noise, *Proc. 1996 Int. Conf. Knowledg Discovery and Data Mining* (*KDD '96*), pp.226–231 (1996).
6) Gionis, A., Mannila, H. and Tsaparas, P.: Clustering aggregation, *Proc. 2005 IEEE Int. Conf. on Data Engineering* (*ICDE '05*), pp.441–352 (2005).
7) Guha, S., Rastogi, R. and Shim, K.: ROCK: A robust clustering algorithm for categorical attributes, *Information Systems*, Vol.25, No.5, pp.345–366 (2000).
8) Hinneburg, A. and Keim, D.A.: Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering, *Proc. 25th Int. Conf. on Very Large Data Bases* (*VLDB '99*), pp.506–517 (1999).
9) Hinneburg, A. and Keim, D.A.: An efficient approach to clustering in large multimedia databases with noise, *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining* (*KDD '98*), pp.58–65 (1998).
10) Karypis, G., (Sam) Han, E.-H. and Kumar, V.: CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, *IEEE Computer*, Vol.32, No.8, pp.68–75 (1999).
11) Milenova, B.L. and Campos, M.M.: O-Cluter: Scalable clutering of large high dimensional data sets, *Proc. 2002 IEEE Int. Conf. on Data Mining* (*ICDM '02*), pp.290–297 (2002).
12) Milenova, B.L. and Campos, M.M.: Clustering large databases with numeric and nominal values using orthogonal projections, *Proc. 29th Int. Conf. on Very Large Data Bases* (*VLDB '03*) (2003).
13) Wang, W., Yang, J. and Muntz, R.: STING: A statistical information grid approach to spatial data mining, *Proc. 23rd Int. Conf. Very Large Data Bases* (*VLDB '97*), pp.186–195 (1997).
14) The University of California, Irvine Knowledge Discovery in Databases Archive: The insurance company benchmark (COIL 2000). http://kdd.ics.uci.edu/
15) 中村朋健，上土井陽子，若林真一，吉田典可：FlexDice: 高次元な大規模データセットに対する高速クラスタリング手法，情報処理学会論文誌：データベース，Vol.46, No.30(TOD 28), pp.40–49 (2005).
16) 中村朋健，上土井陽子，若林真一，吉田典可：FlexDice を用いたクラスタリング結果の特徴抽出，第 16 回データ工学ワークショップ（DEWS '05）論文集，3C-o1 (2005).

(平成 18 年 6 月 21 日受付)
(平成 18 年 10 月 2 日採録)

(担当編集委員　佐藤 哲司)

中村 朋健（学生会員）

2002 年広島市立大学情報科学部卒業．2004 年同大学大学院情報科学研究科博士前期課程修了．現在，同大学院情報科学研究科博士後期課程在学中．

上土井陽子

1989 年広島工業大学電子工学科卒業．1991 年広島大学大学院工学研究科博士課程前期修了．1994 年同大学院工学研究科博士課程後期修了．博士（工学）．1994 年より広島市立大学情報科学部助手．

若林 真一（正会員）

1979 年広島大学工学部電気工学科卒業．1984 年同大学大学院工学研究科博士課程後期修了．同年日本アイ・ビー・エム（株）入社．東京基礎研究所副主任研究員．1988 年広島大学工学部助教授．2003 年より広島市立大学情報科学部教授．工学博士．主として，VLSI CAD，VLSI 設計，組合せ最適化，遺伝的アルゴリズムに関する研究に従事．電子情報通信学会，IEEE，ACM 各会員．

吉田 典可（正会員）

1955 年九州大学工学部通信工学科卒業．1956 年同大学工学部通信工学科助手．同大学工学部電子工学科講師，助教授を経て，1969 年広島大学工学部電子工学科教授（電子回路工学）．1995 年同大学退官，同年広島市立大学情報科学部情報工学科教授（論理回路学）．2003 年同大学退職．工学博士（九州大学）．この間，電子回路とディジタルシステム，論理回路とそのシステム，コンピュータハードウェア，ネットワークシステム等の教育研究に従事．