DEWS2007 E9-6

自然なクラスタリングを選択するための指標関数について

中村　朋健† 　上土井陽子†† 　若林　真一†† 　吉田　典可†††

† 広島市立大学大学院 情報科学研究科　〒 731–3194 広島市安佐南区大塚東三丁目 4–1
†† 広島市立大学 情報科学部　〒 731–3194 広島市安佐南区大塚東三丁目 4–1
††† 前所属 広島市立大学 情報科学部　〒 731–3194 広島市安佐南区大塚東三丁目 4–1
E-mail: †tomotake@ lcl.ce.hiroshima-cu.ac.jp, ††{yoko,wakaba}@ce.hiroshima-cu.ac.jp

あらまし　我々の目標は大規模高次元データをユーザフレンドリにクラスタリング可能にすることである．大規模高次元データでは，ユーザが複数の自然なクラスタリング結果から欲する結果を選択することは困難である．そこで，我々は大規模高次元データをクラスタリングするためには対話的にクラスタリングすることが必要であると考えた．我々は，対話的なクラスタリングを実現するために，1 つのデータセットに対する 2 つのクラスタリング結果において，より自然なクラスタがどちらであるかをユーザに示す選択指標関数を既に提案している．本稿では，外れ度合い算出手法を用い，クラスタリング結果の各クラスタに特異さの度合いを算出する手法と得られた度合いを利用して，より自然なクラスタリング結果を選択可能な選択指標関数を提案する．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選択指標関数の有効性を示す．

キーワード　データマイニング，クラスタリング，高次元データ，選択指標

On Indicator Functions for Evaluating Naturality of Clustering Results

Tomotake NAKAMURA†, Yoko KAMIDOI††, Shin'ichi WAKABAYASHI††, and Noriyoshi YOSHIDA†††

† Graduate School of Information Sciences, Hiroshima City University
3–4–1, Ozuka-higashi, Asaminami-ku, Hiroshima, 731–3194 Japan
†† Faculty of Information Sciences, Hiroshima City University
3–4–1, Ozuka-higashi, Asaminami-ku, Hiroshima, 731–3194 Japan
††† Formerly, Faculty of Information Sciences, Hiroshima City University
3–4–1, Ozuka-higashi, Asaminami-ku, Hiroshima, 731–3194 Japan
E-mail: †tomotake@ lcl.ce.hiroshima-cu.ac.jp, ††{yoko,wakaba}@ce.hiroshima-cu.ac.jp

Abstract　Our aim is to realize user-friendly clustering for large high-dimensional data sets. It is difficult to choose a desired clustering result among natural ones. In order to obtain a desired clustering result for large high-dimensional data sets, we found that it is necessary to perform interactive clustering. We have already proposed an indicator function for choosing a more natural clustering result than the other one. In this paper, we propose a new indicator function based on the degree of outliers, which chooses more natural clustering result than one chosen by the previous indicator function. In simulation experiments with benchmark data, we show effectiveness of the new indicator function.

Key words　Data Mining, Clustering, High dimensional data sets, Indicator

1. はじめに

大規模高次元データセットからユーザにとって有用な情報を効率よく抽出するために，多くのデータマイニングツールが開発されている [4], [7]．これら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セットから簡潔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 [1]．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，つまり類似したデータ要素の集合を見つけ出せる．一般に，クラスタリングによって抽出されたデータは概要を捉えており扱いやすくなることが多く，クラスタリング結果がユーザへの直接的で有益な知識となる．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．入力データの次元数に応じて，妥当なクラスタリング結果の総数が指数関数的に増大するため，クラスタリング結果を効率よく利用することは困難である．

文献 [11] において我々が提案した特徴抽出手法 FEM によって，クラスタリング手法が自然なクラスタリング結果を出力しているかどうか判断することを可能にし，ユーザにクラスタリング結果の特徴を与えることを可能にした．特徴抽出手法 FEM はクラスタリング結果における各クラスタのデータ要素の分布を各属性に投影し，特異な分布を持つクラスタを抽出する手法である．特徴抽出手法 FEM は対話的クラスタリング手法，選択指標関数，そして半教師付きクラスタリング手法 [8] などに応用可能である．対話的クラスタリング手法は文献 [13] において我々が提案した手法であり，ユーザが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手法である．選択指標関数は対話的クラスタリング手法の重要な構成要素であり，2 つのクラスタリング結果から一般的な観点でより自然なクラスタリング結果をユーザに提示することを目標とした手法である．ここで，2 つのクラスタリング結果はクラスタリング手法に 1 つのデータセットと 2 つのパラメータ集合を入力したときに得られる 2 つの結果である．

本稿の目標は，特徴抽出手法 FEM の目標と同じく，複数のクラスタリング結果から欲するクラスタリング結果を選択し易くするために，クラスタリング結果の理解を容易にする情報をユーザに与えることである．特徴抽出手法 FEM は分布算出手法と外れ要素検出手法によって構成される．従来の我々が提案した特徴抽出手法 FEM では外れ要素検出手法に FlexDice [12] を使用していた．本稿では，外れ要素検出手法に外れ度合いを算出可能な手法 [10] を使用することで，クラスタリング結果の自然度合いを算出可能にし，クラスタリング結果の理解を容易にすることを目的とする．具体的には，外れ度合いを算出可能な手法を用いた特徴抽出手法により，特徴抽出手法の応用である選択指標関数と対話的クラスタリング手法の精度の向上を目指す．

本稿では，3. 章において，我々が提案したクラスタリング結果の理解を容易にする情報をユーザに与えることが可能な特徴抽出手法 FEM を説明する．4. 章において，外れ度合い算出手法を説明し，我々が提案する応用例を示す．5. 章において，特徴抽出手法の応用である選択指標関数を提案する．6. 章では，提案した選択指標関数の有効性を示す．最後に 7. 章で，本稿をまとめる．

2. 定　　義

本章において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への入力はオリジナルデータセット ODS であり，クラスタリング手法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 $(1 \leqq i \leqq N_\mathcal{C})$ はオリジナルデータセット ODS 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 の一般的な性質と異なるデータ要素，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．

3. 特徴抽出手法

本章では，クラスタリング結果の理解を容易にする情報をユーザに与えることが可能な特徴抽出手法 FEM(Feature Extraction Method) [13] を説明する．クラスタリング結果の理解を容易にするために，特徴抽出手法 FEM を用いることによって，クラスタリング手法が自然なクラスタリング結果を出力しているかどうか判断することが可能になり，ユーザにクラスタリング結果の特徴を与えることが可能になる．特徴抽出手法 FEM はクラスタリング結果における各クラスタのデータ要素の分布を各属性に投影し，特異な分布を持つクラスタを抽出する手法である．特徴抽出手法 FEM はクラスタリング結果の特徴として属性ごとに特異な分布を持つクラスタとそのクラスタのデータ要素の分布を出力する．

3.1 特徴抽出手法

図 1 に特徴抽出手法 FEM の処理の流れを示す．特徴抽出手法 FEM は分布算出手法（DCM: Distribution Computing Method）と外れ要素検出手法（ODM: Outlier Detection Method）から構成される．外れ要素検出手法とはデータセットから特異な要素を検出する手法である．特徴抽出手法は 5.1 節で説明する選択指標関数にも使用され，拡張した特徴抽出手法は 5.2 節で使用される．以降では，特徴抽出手法 FEM のアルゴリズムとその重要な構成要素である分布算出手法 DCM について説明する．

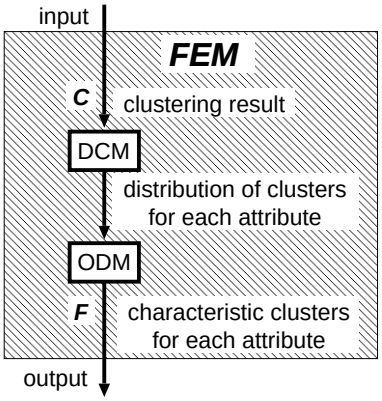

図 1　特徴抽出手法 FEM の流れ

Algorithm *ODM*(*Original data Set*:*ODS*, e_i, layer);
begin
  for $j = 1$ **to** *layer* **do**
  begin
    $r_j = ComputeLength(j, ODS)$;
    $E_j = EnumerateCells(r_j, ODS)$;
    $E'_j = EnumerateCells(r_j, ODS - \{e_i\})$;
  end;
  $Fd = ComputeFD(\{(r_j, E_j) \mid 1 \leqq j \leqq layer\})$;
  $Fd' = ComputeFD(\{(r_j, E'_j) \mid 1 \leqq j \leqq layer\})$;
  return$(Fd' - Fd)$;
end

図 2　外れ度合い算出手法

特徴抽出手法 FEM への入力はクラスタリング手法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．ここで，クラスタ集合 $\mathcal{C}'$ はクラスタリング結果 $\mathcal{C}$ からノイズクラスタを除いたクラスタ集合である．クラスタリング結果 $\mathcal{C}$ の特徴 $\mathcal{F}$ は，各クラスタが特異な分布を持つ属性と，その属性におけるデータ要素の分布で表される．

分布算出手法 DCM の入力は特徴抽出手法 FEM の入力と同じクラスタリング結果，つまりクラスタ集合 $\mathcal{C}' = \{C_1, \ldots, C_{N_\mathcal{C}}\}$ である．分布算出手法 DCM の出力であり外れ要素検出手法 ODM の入力は，各属性 d（$1 \leqq d \leqq dim$）に関する各クラスタ C_i の分布を表すベクトル集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leqq i \leqq N_\mathcal{C}\}$ である．$\boldsymbol{V}(C_i, d)$ をクラスタベクトルと呼ぶ．分布算出手法の入力データの属性 d の値域は 1 以上 $N^d_{\max}$ 以下の自然数の部分集合であると仮定する．ここで，$N^d_{\max}$ は自然数の定数とする．属性値が実数である場合や，値域が広すぎる場合は離散化して自然数の値を割り当てる．このように変換した各属性の値を値識別子（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 (1) で表される．

$$\boldsymbol{V}(C_i, d) = \frac{100}{N(C_i)}(v(i, d, 1), \ldots, v(i, d, m(d))) \qquad (1)$$

ここで $N(DS)$ は集合 DS 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．

外れ要素検出手法 ODM は，属性 d におけるクラスタベクトルの集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leqq i \leqq N(\mathcal{C})\}$ 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（$1 \leqq d \leqq dim$）に関するクラスタベクトルの集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leqq i \leqq N(\mathcal{C})\}$ を外れ要素検出手法 ODM に入力することで，類似するクラスタベクトルが少ないクラスタベクトルの集合を抽出できる．したがって，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を特徴抽出手法 FEM は抽出する．外れ要素検出手法 ODM として，上記の条件を満たしてデータセットから外れ要素を検出可能な手法として LOF [3]，DBSCAN [4]，FlexDice [12]，そして伏見らによる外れ要素検出手法 [10] などが挙げられる．

3.2　特徴抽出手法の利用法と拡張

データセットには自然なクラスタリング結果が属性数に対して指数関数的に増加する可能性があるので，ユーザは質の高いクラスタリング手法を利用しても欲する結果を得ることが困難なことがある．特徴抽出手法の目標はクラスタリング結果の理解を容易にする情報をユーザに与え，より自然なクラスタリング結果をユーザが得やすくすることである．特徴抽出手法は各属性に関して特異なクラスタを抽出可能である．特異なクラスタを多く含む結果は自然なクラスタリング結果であると推測できるため，特徴抽出手法によって自然なクラスタリング結果を選択しやすくなる．

特徴抽出手法によって得られる特異な分布を持つクラスタに特異さの度合いを付けることによって，我々はクラスタリング結果の自然な度合いをユーザに提示可能になると考えた．得られたクラスタリング結果におけるクラスタに特異さの度合いを付けるために，外れ要素検出手法を外れ度合い検出手法にした．この拡張によって，より自然なクラスタリング結果をユーザに提示すること目指す．

4.　外れ度合い算出

4.1　外れ度合い算出手法

提案選択指標関数において外れ度合い算出手法を使用する．本節では，6. 章以降のシミュレーション実験において使用する外れ度合い算出手法 [10] を説明する．外れ度合い算出手法 [10] ではフラクタル次元を利用してストリームデータのデータ要素の外れ度合いを算出している．上記手法ではフラクタル次元算出手法として box-counting 法を用いている．

図 2 に外れ度算出手法 *ODM* を示す．外れ度算出手法への入力はオリジナルデータセット *ODS*，外れ度合いを調べたいデータ要素 e_i，そして階層数 *layer* である（$ODS = \{e_i \mid 1 \leqq i \leqq N\}$）．

データセット *ODS* の多次元データ空間を一辺の長さが r_j の

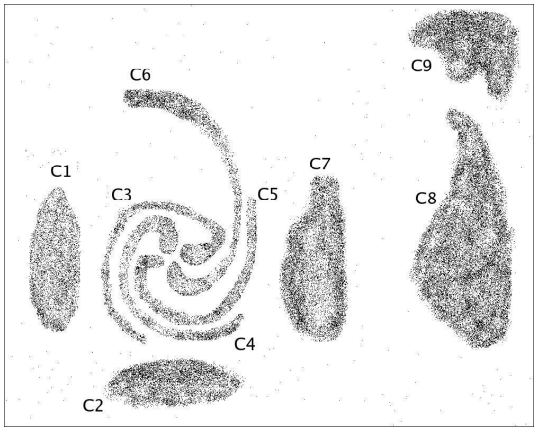

図 3　入力データセット

超立方体（セル）に分割する．データ要素が含まれているセルの数を E_j とする．データセット $ODS-\{e_j\}$ の多次元データ空間を一辺の長さが r_j の超立方体（セル）に分割したときにおいて，データ要素が含まれているセルの数を E'_j とする．各 $j(1 \leqq j \leqq layer)$ 階層のセルの一辺の長さ r_j に関する 2 次元座標上の $layer$ 個の点 (r_j, E_j) と点 (r_j, E'_j) を算出する．次に，$layer$ 個の点 (r_j, E_j) と点 (r_j, E'_j) に対して，それぞれフラクタル次元 Fd,Fd' を算出する．ここで，フラクタル次元とは，フラクタルの度合いを定量的に表す次元である．ただし，フラクタル次元は整数とは限らない．一般に，フラクタル次元が大きいとそのデータセットは複雑な分布をしており，逆に，フラクタル次元が小さいとそのデータセットは単純な分布をしている．$layer$ 個の点 $(r_j, E_j)(1 \leqq j \leqq layer)$ のフラクタル次元 Fd は，セルの一辺の長さ r_j の対数 $\log r_j$ とデータ要素を含むセルの数 E_j の対数 $\log E_j$ の関係の傾きである．この傾き（フラクタル次元）は最小二乗法で算出する．$layer$ 個の点 (r_j, E'_j) のフラクタル次元 Fd' も同様に算出する．最後に，データ要素 e_i の外れ度合いを $|Fd' - Fd|$ として算出する．

4.2　クラスタベクトルの外れ度合い算出例

本節において，図 3 のデータセットに対するクラスタリング結果における各属性のクラスタベクトルの外れ度合い算出例を示す．横軸が属性 x，縦軸が属性 y であるデータ空間上にデータ要素が散らばっている．データ要素はデータ空間上の黒い点である．図 3 をクラスタリングするとクラスタ集合 $\{C_i \mid 1 \leqq i \leqq 9\}$ が形成された．

クラスタ集合 $\{C_i \mid 1 \leqq i \leqq 9\}$ の属性 x，属性 y に関するクラスタベクトルを図 4，図 5 に示す．属性 x に関しては，25 次元のクラスタベクトルのフラクタル次元を求めることで，外れ度合いを算出する．属性 x に関する外れ度合い算出手法への入力は，図 4 に示したすべてのクラスタベクトル $\boldsymbol{V}(C_i, x)(1 \leqq i \leqq 9)$，外れ度合いを算出したいクラスタベクトル $\boldsymbol{V}(C, x)$，そして階層数 $layer$ である．属性 y も同様に，18 次元のクラスタベクトルの外れ度合いを算出する．属性 y に関する外れ度合い算出手法への入力は，属性 x と同様である．表 1 に，文献 [10] の手法による各属性に関する各クラスタベクトルの外れ度合いの算出結果を示す．

表 1　各クラスタベクトルの外れ度合い

	$attribute$	
	x	y
$\boldsymbol{V}(C_1, attribute)$	0.770	0.596
$\boldsymbol{V}(C_2, attribute)$	0.692	0.651
$\boldsymbol{V}(C_3, attribute)$	0.692	0.596
$\boldsymbol{V}(C_4, attribute)$	0.692	0.596
$\boldsymbol{V}(C_5, attribute)$	0.692	0.596
$\boldsymbol{V}(C_6, attribute)$	0.692	0.651
$\boldsymbol{V}(C_7, attribute)$	0.770	0.596
$\boldsymbol{V}(C_8, attribute)$	0.678	0.581
$\boldsymbol{V}(C_9, attribute)$	0.678	0.651

表 1 の結果より，属性 x に関してクラスタ C_1 と C_7 に対応するクラスタベクトル $\boldsymbol{V}(C_1, x)$ と $\boldsymbol{V}(C_7, x)$ が最も外れ度合いが高いことが分かる．属性 y に関してクラスタ C_2 と C_6 に対応するクラスタベクトル $\boldsymbol{V}(C_2, y)$ と $\boldsymbol{V}(C_6, y)$ が最も外れ度合いが高いことが分かる．

この外れ度合い算出手法を特徴抽出手法 FEM の外れ要素検出手法に使用するためには，外れ度合いがある値以上であるクラスタベクトルを外れ要素とする．例えば属性 x に関して 0.68 以上であるクラスタベクトルを外れ要素とするときは，クラスタベクトル $\boldsymbol{V}(C_1, x)$，$\boldsymbol{V}(C_2, x)$，$\boldsymbol{V}(C_3, x)$，$\boldsymbol{V}(C_4, x)$，$\boldsymbol{V}(C_5, x)$，$\boldsymbol{V}(C_6, x)$，$\boldsymbol{V}(C_7, x)$ が外れているクラスタベクトルとなる．属性 y に関して 0.60 以上であるクラスタベクトルを外れ要素とするときは，クラスタベクトル $\boldsymbol{V}(C_2, y)$，$\boldsymbol{V}(C_6, y)$，$\boldsymbol{V}(C_9, y)$ が外れているクラスタベクトルとなる．

5.　選択指標関数

クラスタリング手法によって出力された 2 つのクラスタリング結果に対し，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くラスタに集め，類似していないデータ要素を別々のクラスタに分け，そして各クラスタが特異な分布を持つクラスタリングのことと定義する．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 を説明する．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．

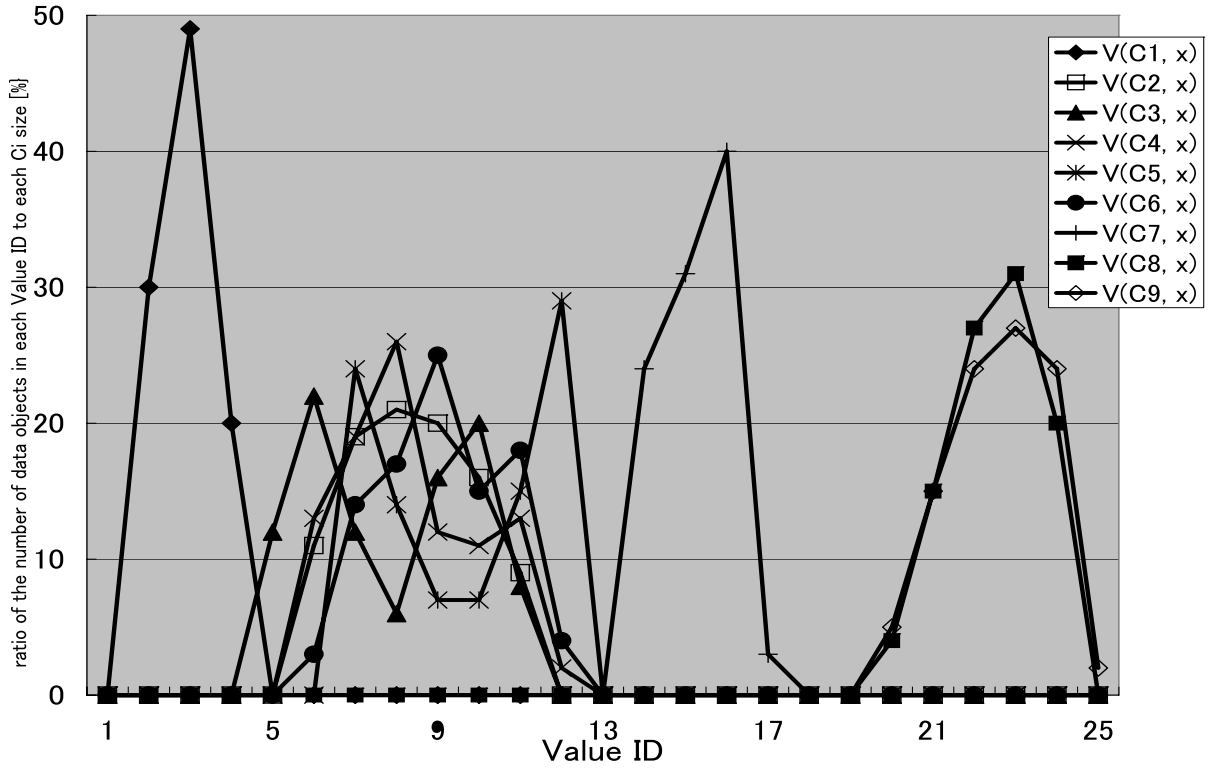

図 4　各クラスタに対応した属性 x における各クラスタベクトル

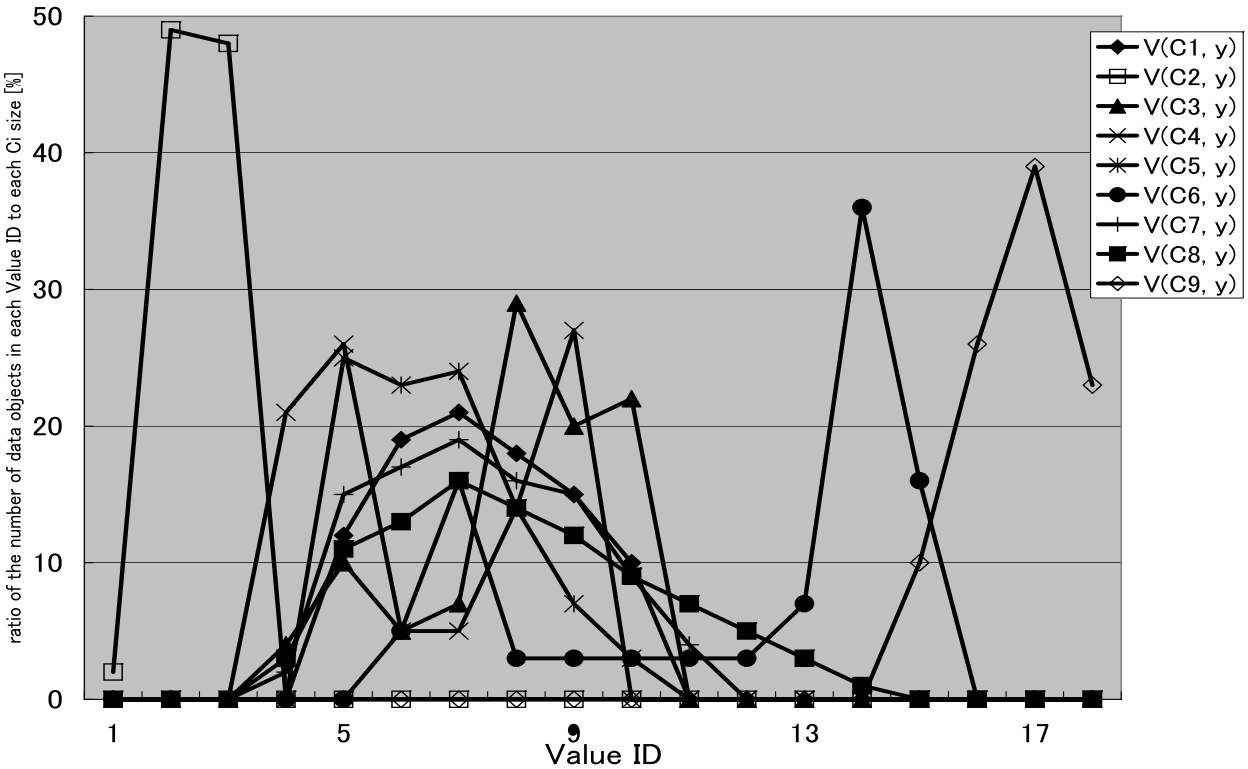

図 5　各クラスタに対応した属性 y における各クラスタベクトル

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 $\mathcal{C}_A = \{C_{A1}, \ldots, C_{AN_C}, Noise_{\mathcal{C}_A}\}$, $\mathcal{C}_B = \{C_{B1}, \ldots, C_{BN_C}, Noise_{\mathcal{C}_B}\}$ に対応するノイズクラスタを除いたクラスタ集合 $\mathcal{C}'_A = \{C_{A1}, \ldots, C_{AN_C}\}$, $\mathcal{C}'_B = \{C_{B1}, \ldots, C_{BN_C}\}$ 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，ユーザはどちらが良いのか判断することが要求される．

5.1 従来の選択指標関数

以下で，与えられた 2 つのクラスタリング結果 $\mathcal{C}_A = \{C_{A1}, \ldots, C_{AN_{\mathcal{C}_A}}, Noise_{\mathcal{C}_A}\}$, $\mathcal{C}_B = \{C_{B1}, \ldots, C_{BN_{\mathcal{C}_B}}, Noise_{\mathcal{C}_B}\}$ に対する従来の選択指標 I の決定方法を示す．

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 $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))(1 \leqq l \leqq L)$ 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．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，$\alpha, \beta, \gamma, \delta$ 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する．

上記の選択指標値算出の概要おいて，2 つのクラスタリング結果のクラスタ集合の対応付けについて説明する．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も特徴であるかどうか調べる．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結果間での対応を持つようにそれぞれ分割し，クラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ を構成する．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（2）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．

$$\bigcup_{C_X \in \mathcal{S}_l(\mathcal{C}'_A)} C_X - Noise_{\mathcal{C}_B} = \bigcup_{C_Y \in \mathcal{S}_l(\mathcal{C}'_B)} C_Y - Noise_{\mathcal{C}_A} \quad (2)$$

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．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は文献 [13] を参照されたい．

選択指標 I の値は先に定義した $\alpha, \beta, \gamma, \delta$ の値によって決定する．$\alpha, \beta, \gamma, \delta$ の中で α が最も大きいとき $I = 1$，β が最も大きいとき $I = 2$，それ以外のとき $I = 3$ となる．

5.2 提案選択指標関数

提案する選択指標関数も 5.1 節で説明した従来の選択指標関数と同様に 2 つのクラスタリング結果において，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することを目的とする．提案する選択指標関数は従来の選択指標関数の処理である “L 個のクラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))(1 \leqq l \leqq L)$ を算出する” まで同様な処理をする．その後，従来の選択指標関数と提案選択指標関数はクラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{S}_l(\mathcal{C}'_B)$ のどちらが興味深いかを調べる．従来の選択指標関数はクラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{S}_l(\mathcal{C}'_B)$ に関して，より自然なクラスタリングを特徴抽出手法で求める．一方で，提案選択指標関数はクラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$ と $\mathcal{S}_l(\mathcal{C}'_B)$ に関して，より自然なクラスタリングを外れ度合い算出手法で求める．

以下で説明する選択指標関数の変数や集合の表記は 2. 章，3.1 節，5.1 節と同様の表記を使用する．与えられた 2 つのクラスタリング結果 $\mathcal{C}_A$，$\mathcal{C}_B$ に対する選択指標 I の決定方法を図 6 に示す．

Step 3 の外れ度合い算出手法には LOF や文献 [10] で提案された手法を用いる．***Step 4*** における，クラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ から自然なクラスタが形成されているクラスタ集合を選択する方法を以下で説明する．d は外れ度合いを求めたい属性であり，N_{vec} は入力パラメータである．$AveOut(\mathcal{S}_l(\mathcal{C}'_A), d, N_{vec})$ を $\mathcal{S}_l(\mathcal{C}'_A)$ に含まれるクラスタに対応する属性 d に関する外れ度合いが高い N_{vec} 個のクラスタベクトルの属性 d に関する外れ度合いの平均とする．このとき，もしクラスタ集合 $\mathcal{S}_l(\mathcal{C}'_A)$ の要素が N_{vec} 以下の場合，すべてのクラスタベクトルの属性 d に関する外れ度合いの平均を $AveOut(\mathcal{S}_l(\mathcal{C}'_A), d, N_{vec})$ とする．同様に $AveOut(\mathcal{S}_l(\mathcal{C}'_B), d, N_{vec})$ によってクラスタ集合 $\mathcal{S}_l(\mathcal{C}'_B)$ から選択したクラスタベクトルの属性 d に関する外れ度合いの平均を表す．式 (3) が成り立つとき，α をインクリメントする．

Step 1: 2 つのクラスタリング結果 $\mathcal{C}_A$，$\mathcal{C}_B$ に対応する $\mathcal{C}'_A$，$\mathcal{C}'_B$ におけるクラスタの要素の対応関係を調べる．$\mathcal{C}'_A$ 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 $(1 \leqq l \leqq L)$.

Step 2: 3.1 節で説明した分布算出手法を用いて，各属性 d（$1 \leqq d \leqq dim$）に関する各クラスタ C_i の分布を表すベクトル集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leqq i \leqq N_{\mathcal{C}}\}$ を算出する．

Step 3: データセットにおけるデータ要素の外れ度合いを算出可能な手法（外れ度合い算出手法）を用いて，各ベクトル集合 $VD(d) = \{\boldsymbol{V}(C_i, d) \mid C_i \in \mathcal{C}', 1 \leqq i \leqq N_{\mathcal{C}}\}$ における，各ベクトル $\boldsymbol{V}(C_i, d)$ の外れ度合い $ODM(VD(d), \boldsymbol{V}(C_i, d), layer)$ を算出する．ここで，$layer$ は外れ度合い算出手法の入力パラメータである．

Step 4: 対応する L 個のクラスタ集合のペア $(\mathcal{S}_l(\mathcal{C}'_A), \mathcal{S}_l(\mathcal{C}'_B))$ から自然なクラスタが形成されているクラスタ集合を選択する．クラスタリング結果 $\mathcal{C}_A$ に対応したクラスタ集合を選択した数を α，クラスタリング結果 $\mathcal{C}_B$ に対応したクラスタ集合を選択した数を β とする．$\alpha > \beta$ のとき $I = 1$，$\alpha < \beta$ のとき $I = 2$，$\alpha = \beta$ のとき $I = 3$ とする．

図 6　提案選択指標関数

$$
\begin{aligned}
&|\{d : AveOut(\mathcal{S}_l(\mathcal{C}'_A), d, N_{vec}) > \\
&\qquad AveOut(\mathcal{S}_l(\mathcal{C}'_B), d, N_{vec})\}| > \\
&|\{d : AveOut(\mathcal{S}_l(\mathcal{C}'_A), d, N_{vec}) < \\
&\qquad AveOut(\mathcal{S}_l(\mathcal{C}'_B), d, N_{vec})\}| \qquad (3)
\end{aligned}
$$

式 (4) が成り立つとき，β をインクリメントする．

$$
\begin{aligned}
&|\{d : AveOut(\mathcal{S}_l(\mathcal{C}'_A), d, N_{vec}) > \\
&\qquad AveOut(\mathcal{S}_l(\mathcal{C}'_B), d, N_{vec})\}| < \\
&|\{d : AveOut(\mathcal{S}_l(\mathcal{C}'_A), d, N_{vec}) < \\
&\qquad AveOut(\mathcal{S}_l(\mathcal{C}'_B), d, N_{vec})\}| \qquad (4)
\end{aligned}
$$

以上により，α, β の値を求め，図 6 の ***Step 4*** により選択指標を求める．

6. シミュレーション実験

我々が提案する選択指標関数は対話的に自然なクラスタリング結果を得るために手法である．本章のシミュレーション実験では提案選択指標関数を用いて対話的にクラスタリング結果を得たとき，応用の一例として考えられる分類手法における精度を調べる．

6.1 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．文献 [5] などでは，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使用される分類エラー E_C が使用されている．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たことを高く評価し難い．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エラー $E'_{\mathcal{C}}$ を文献 [13] で定義した．クラスタリングエラー $E'_{\mathcal{C}}$ 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．

実験において使用するベンチマークデータは UCI KDD アーカイブ [9] からのデータである．収入データは分類に関するデータであり 14 属性を含む 32,561 個のデータ要素が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する．

以降の実験において，選択指標関数に与える入力パラメータ $layer$ は 3 とした．

6.2 提案選択指標関数の分類手法への応用

本稿で提案した選択指標関数は，2 つのクラスタリング結果から 1 つの結果を選択するための指標を示すことができた回数は，15 回の試行のうち 14 回であった．提案した選択指標関数は 2 つのクラスタリング結果から自然なクラスタリング結果を選択し，選択したクラスタリング結果が他方のクラスタリング結果よりもクラスタリングエラー $E'_{\mathcal{C}}$ が少ない場合は 14 回の試行において 13 回であった．この結果より，多くの場合において提案選択指標関数はクラスタリングエラーの少ない結果を自然なクラスタリング結果としてユーザに提示できることがわかった．

6.3 対話的クラスタリング手法の実行例

提案した選択指標関数を組み込んだ対話的クラスタリングを実行した例を示す．入力は収入データであり，クラスタリング手法には FlexDice [12] を使用した．ユーザが選択した結果はすべて選択指標関数が提示する結果を選択した．表 2 に FlexDice へ入力したパラメータとそのときのクラスタリングエラー $E'_{\mathcal{C}}$

表 2　各結果における入力パラメータ値と $E'_{\mathcal{C}}$ の値

	P_d	P_b	clusters	$E'_{\mathcal{C}}$
$\mathcal{C}_A$	10	2	3	0.499
$\mathcal{C}_B$	200	2	12	0.462
$\mathcal{C}_C$	300	2	18	0.356
$\mathcal{C}_D$	300	3	126	0.296
$\mathcal{C}_E$	300	4	129	0.382

を示す．ここで，P_d は子セル数が P_d 以上であるならば，分割を進めないセルとし，P_b は最下位層数である．本実験において，FlexDice の他の入力パラメータ P_{min} は 1，P_{ele} は 10 を入力した．ここで P_{min} はデータ要素数によってセル無いのデータ要素を外れ要素とするのかしないのかを決定するパラメータであり，P_{ele} はクラスタに含まれるデータ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．

以下に 2 つクラスタリング結果を提案選択指標関数に入力したときの選択した結果を示す．

（1） 結果 $\mathcal{C}_A$ と結果 $\mathcal{C}_B$ を入力したとき，結果 $\mathcal{C}_B$ を選択
（2） 結果 $\mathcal{C}_B$ と結果 $\mathcal{C}_C$ を入力したとき，結果 $\mathcal{C}_C$ を選択
（3） 結果 $\mathcal{C}_C$ と結果 $\mathcal{C}_D$ を入力したとき，結果 $\mathcal{C}_D$ を選択
（4） 結果 $\mathcal{C}_D$ と結果 $\mathcal{C}_E$ を入力したとき，結果 $\mathcal{C}_D$ を選択

以上のように，ユーザとクラスタリング手法の 4 度の対話によりクラスタリングエラー $E'_{\mathcal{C}}$ を 0.499 から 0.296 まで減少させることができた．提案選択指標関数が対話的にクラスタリングエラーの小さい結果を選択できたことを示した．分類手法から得られる 2 つの結果を対話的に選択することで，精度の高い分類結果をユーザが得得られることが予測できる．

7. おわりに

本稿では，複数のクラスタリング結果から欲するクラスタリング結果を選択し易くするために，クラスタリング結果の理解を容易にする情報をユーザに与えることを目指した．本稿で提案した選択指標関数を分類手法に応用したとき，エラーの少ない結果を選択可能であることをシミュレーション実験により示した．今後の課題として，提案選択指標関数が自然なクラスタリング結果を対話的に得られているかどうか分析することが挙げられる．

文　献

[1] C. C. Aggarwal, “Towards effective and interpretable data mining by visual interaction,” ACM-SIGKDD Explorations, Volume 3, pp. 11–22, 2002.

[2] C. C. Aggarwal, “Towards exploratory test instance specific algorithms for high dimensional classification,” Proc. of the 11st ACM-SIGKDD Int. Conf. on Knowledge discovery in data mining (KDD ’05), pp. 526–531, 2005.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Hg and J. Sander, “LOF: Identifying density-based local outliers,” Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD ’00), pp. 93–104, 2000.

[4] M. Ester, H. -P. Kriegel, J. Sander and X. Xu, “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.

[5] A. Gionis, H. Mannila and P. Tsaparas, “Clustering aggregation,” Proc. 2005 IEEE Int. Conf. on Data Engineering (ICDE ’05), pp. 441–352, 2005.

[6] A. Hinneburg and D. A. Keim, “Optimal Grid-Clustering: Towards breaking the curse of dimensionality in high-dimensional clustering,” Proc. of the 25th Int. Conf. on Very Large Data Bases (VLDB ’99), pp. 506–517, 1999.

[7] B. L. Milenova and M. M. Campos, “O-Cluter: Scalable clustering of large high dimensional data sets,” Proc. of the 2nd IEEE Int. Conf. on Data Mining (ICDM ’02), pp. 290–297, 2002.

[8] T. Nakamura, Y. Kamidoi, S. Wakabayashi and N. Yoshida, “A decision method of attribute importance for classification by outlier detection,” IEEE Computer Press. Second Interntional Workshop on Databases for Next-Generation Researchers (SWOD’06), pp.45-50, 2006.

[9] The University of California, Irvine Knowledge Discovery in Databases Archive, “The insurance company benchmark (COIL 2000),” http://kdd.ics.uci.edu/.

[10] 伏見 健史，“データマイニングにおけるストリームデータに対するフラクタル次元を用いた外れ度算出アルゴリズム”, 広島市立大学大学院情報科学研究科情報工学専攻修士論文, 2006.

[11] 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.

[12] 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 (電子情報通信学会 データ工学研究専門委員会共同編集), Vol. 46, No. SIG 18 (TOD 28), pp. 40–49, 2005.

[13] 中村 朋健，上土井 陽子，若林 真一，吉田 典可，“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”, 情報処理学会論文誌: データベース (電子情報通信学会 データ工学研究専門委員会共同編集), Vol. 47, No. SIG 19 (TOD 32), pp. 28–41, 2006.