

# K平均クラスタリングにおけるKの選択

D T Pham\*, S S Dimov, and C D Nguyen

Manufacturing Engineering Centre, Cardiff University, Cardiff, UK

The manuscript was received on 26 May 2004 and was accepted after revision for publication on 27 September 2004.

DOI: 10.1243/095440605X8298

**概要：**K-meansアルゴリズムは、一般的なデータクラスタリングアルゴリズムである。しかし、その欠点の1つは、アルゴリズムを適用する前にクラスター数Kを指定する必要があることである。本稿ではまず、アルゴリズムのクラスタ数を選択するための既存の方法をレビューする。次に、この選択に影響を与える要因について議論し、選択を支援する新しい尺度を提案する。本稿は、異なるデータセットに対してK-meansアルゴリズムのクラスタ数を決定するために提案された尺度を使用した結果の分析で締めくくられる。

**キーワード：**クラスタリング, K平均アルゴリズム, クラスタ数の選択

## 1 INTRODUCTION

データクラスタリングは、類似した特徴を持つオブジェクトをグループ化し、その後の処理を容易にするデータ探索技術である。データクラスタリングは、細胞製造のための部品ファミリーの識別など、多くの工学的応用がある。

K-means アルゴリズムは、一般的なデータクラスタリング・アルゴリズムである。これを使用するには、データのクラスタ数を事前に指定する必要がある。与えられたデータセットに対して適切なクラスター数を見つけることは、一般的に試行錯誤のプロセスであり、何が「正しい」クラスタリングを構成するかを決定する主観的な性質によってより困難になっている[1]。

本論文では、クラスタ数Kを選択するために、K-means クラスタリング操作自体で得られた情報に基づく方法を提案する。この方法は、Kの適切な値を提案するために客観的な評価尺度を採用することで、試行錯誤の必要性を回避する。

本稿の残りの部分は5つのセクションで構成される。セクション3では、Kの選択に影響を与える要因を分析する。セクション5では、異なるデータセットに対してKを選択するために提案された尺度を適用した結果を示す。セクション6は本稿の結論である。

## 2 クラスター数の選択とクラスタリングの妥当性評価

このセクションでは、K-meansアルゴリズムのKを選択するための既存の方法と、それに対応するクラスタリング検証技法についてレビューする。

### 2.1 範囲または集合内で指定されるKの値

クラスタリングアルゴリズムの性能は、選択されたKの値に影響される可能性がある。データセットの特性を反映するために、考慮される値の数が適度に大きいことが重要である。同時に、選択された値は、データセット内のオブジェクトの数よりもかなり小さくなければならない。

K-meansクラスタリングとその応用に関する報告された研究[2-18]は、通常、Kの特定の値を選択するための説明や正当性を含んでいません。表のデータを分析すると、2つの観察ができる。第一に、多くの研究者[5-7, 9]はKに1つか2つの値しか使っていない。第二に、他のいくつかの研究者[1, 3, 11, 13, 16]は、オブジェクトの数に比べて比較的大きなK値を利用している。したがって、クラスタリング結果は、テストされたアルゴリズムの性能を必ずしも正しく表していない。

\*Corresponding author: Manufacturing Engineering Centre, Cardiff University, Cardiff CF24 OYF, UK.

**Table 1** K平均アルゴリズムの異なる研究で使用されたクラスタ数

Reference	Numbers of clusters $K$	Number of objects $N$	Maximum $K/N$ ratio (%)
[2]	32, 64, 128, 256, 512, 1024	8 192	12.50
[3]	32, 64, 128, 256, 512, 1024 256	29 000 2 048	10.00
[4]	600, 700, 800, 900, 1000	10 000	0.13
[5]	600, 700, 800, 900, 1000	50 000	0.70
[6]	4, 16, 64, 100, 128	100 000	
[7]	4, 16, 64, 100, 128	120 000	
[8]	4, 16, 64, 100, 128	256 000	
[9]	4	564	
[10]	4	720	
[11]	4	1 000	
[12]	4	1 008	
[13]	4	1 010	
[14]	4	1 202	
[15]	4	2 000	
[16]	4	2 324	
[17]	4	3 005	
[18]	4	4 000	
[19]	4	6 272	
[20]	4	7 561	
[21]	6	150	4.00
[22]	10	2 310	0.43
[23]	25	12 902	
[24]	2, 4, 8	Not reported	Not reported
[25]	2, 4	500	3.33
[26]	2, 4	50 000	
[27]	2, 4	100 000	
[28]	10	300	
[29]	1, 2, 3, 4	10 000	0.04
[30]	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	500	20.00
[31]	100	10 000	2.00
[32]	50	2 500	
[33]	7	42	16.66
[34]	1, 2, 3, 4, 5, 6, 7	120	
[35]	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	250	5.60
[36]	8, 20, 50, 64, 256	10 000	2.56
[37]	5000	50 000	50.00
[38]	5000	100 000	
[39]	5000	200 000	
[40]	5000	300 000	
[41]	5000	433 208	
[42]	100	100 000	
[43]	250	200 000	
[44]	1000	100 000	
[45]	1000	200 000	
[46]	1000	300 000	
[47]	1000	433 208	
[48]	40	20 000	
[49]	10, 20, 30, 40, 50, 60, 70, 80	30 000	
[50]	50, 500, 5000	10 000	
[51]	50, 500, 5000	50 000	
[52]	50, 500, 5000	100 000	
[53]	50, 500, 5000	200 000	
[54]	50, 500, 5000	300 000	
[55]	50, 500, 5000	433 208	

(continued)

**Table 1** Continued

Reference	Numbers of clusters $K$	Number of objects $N$	Maximum $K/N$ ratio (%)
[17]	250	80 000	10.00
[18]	250	90 000	
[19]	250	100 000	
[20]	250	110 000	
[21]	250	120 000	
[22]	50, 100, 400	4 000	
[23]	50, 100, 400	36 000	
[24]	250	80 000	
[25]	250	90 000	
[26]	250	100 000	
[27]	250	110 000	
[28]	250	120 000	
[29]	50, 100, 150	4 000	
[30]	50, 100, 150	36 000	
[31]	50	800 000	
[32]	500	800 000	
[33]	3, 4	150	6.67
[34]	4, 5	75	
[35]	2, 7, 10	214	

一般的に、K-meansアルゴリズムの新バージョンの性能は、同じ基準で先行バージョンと比較することで検証できる。特に、クラスタ歪みの総和は通常このような性能指標として採用される[3, 6, 13, 16, 18]。このように、性能分析には同じモデルと基準が使用されるため、比較は公平であると考えられる。

## 2.2 ユーザーが指定したKの値

多くのデータマイニングまたはデータ分析ソフトウェアパッケージ[19– 22]におけるK-meansアルゴリズムの実装では、ユーザーがクラスタ数を指定する必要があります。満足のいくクラスタリング結果を見つけるには、通常、ユーザーがKの値を変えてアルゴリズムを実行する反復回数が必要です。このアプローチでは、ユーザーが多次元データセットのクラスタリング結果を評価することは困難である。

## 2.3 後の処理ステップで決定されるKの値

K-meansクラスタリングが前処理ツールとして使用される場合、クラスタ数は主処理アルゴリズムの特定の要件によって決定される[13]。このアルゴリズムのパフォーマンスに対するクラスタリング結果の影響には注意が払われません。このようなアプリケーションでは、K-meansアルゴリズムは、クラスタリング結果の検証なしに、単なる「ブラックボックス」として採用されます。

## 2.4 ジェネレータ数と等しいKの値

アルゴリズムをテストするために使用される合成データセットは、多くの場合、正規分布または一様分布生成器のセットによって作成されます。そして、クラスタリングアルゴリズムは、クラスタの数がジェネレータの数と等しくなるように、これらのデータセットに適用されます。どのような結果のクラスタも、特定のジェネレータによって作成されたすべてのオブジェクトをカバーすると仮定されます。したがって、クラスタリングの性能は、クラスタによってカバーされるオブジェクトと、対応するジェネレータによって作成されたオブジェクトとの間の差に基づいて判断される。このような差は、単純にオブジェクトを数えるか、情報利得を計算することで測定することができます[7]。

この方法には欠点がある。最初の欠点は、オブジェクト空間内に異なるジェネレーターによって作成されたオブジェクトを含む領域がある場合のクラスタリング結果の安定性に関するものである。図1aはそのようなケースを示している。この図に示すデータセットにはAとBの2つのクラスタがあり、それぞれ生成子 $G_A$ と $G_B$ によって生成されたオブジェクトをカバーしている。

オブジェクトXはクラスタAとBの重複領域にある。Xはそれぞれ $G_A$ と $G_B$ によって生成される確率 $P_{GA}$ と $P_{GB}$ を持ち、それぞれクラスタAとBに含まれる確率 $P_{CA}$ と $P_{CB}$ を持つ。したがって、Xが生成元 $G_A$ によって生成されてもクラスタBでカバーされる可能性があり、その逆もある。このような場合、クラスタリング結果は完全ではない。クラスタリング結果の安定性はこれら4つの確率に依存する。オブジェクト空間の重複領域が増加すると、クラスタリング結果の安定性は低下する。

ジェネレーターの特性の違いもクラスタリングの結果に影響を与える。クラスタAのオブジェクト数がクラスタBのオブジェクト数の5倍である図1bでは、小さいクラスタBはノイズとみなされ、すべてのオブジェクトが1つのクラスタにグループ化されるかもしれない。このようなクラスタリングの結果は、目視で得られたものとは異なる。

残念ながら、このKの選択方法は実用的な問題には適用できない。現実的な問題でのデータ分布は未知であり、またジェネレータの数も指定できない。

## 2.5 統計的尺度によって決定されるKの値

これらの尺度は、確率的クラスタリングアプローチと組み合わせて適用されることが多い。これらの統計的尺度は、データの基礎となる分布についてある仮定において計算されます。ベイズ情報量規準またはAkeikeの情報量規準[14, 17]は、ガウス分布の集合で構成されるデータセットで計算されます。Hardy [23]が適用した尺度は、データセットがポアソン分布に適合するという仮定に基づいています。帰無仮説に関連するモンテカルロ技法は、クラスタリングの結果を評価し、またクラスタ数を決定するために使用されます[24, 25]。

確率的クラスタリングと分割クラスタリングの比較がある [7]。期待値最大化(EM)は確率的クラスタリングの代表的な手法としてよく知られています。同様に、K-meansクラスタリングはパーティショニング・クラスタリングの代表的な手法と考えられています。EMとK-meansクラスタリングには共通の考え方がありますが、異なる仮説、モデル、基準に基づいています。確率的クラスタリング手法はクラスタ内部の歪みを考慮しないため、このような手法を適用して作成されたクラスタは分割クラスタリングにおけるクラスタに対応しない可能性があり、またその逆も同様である。

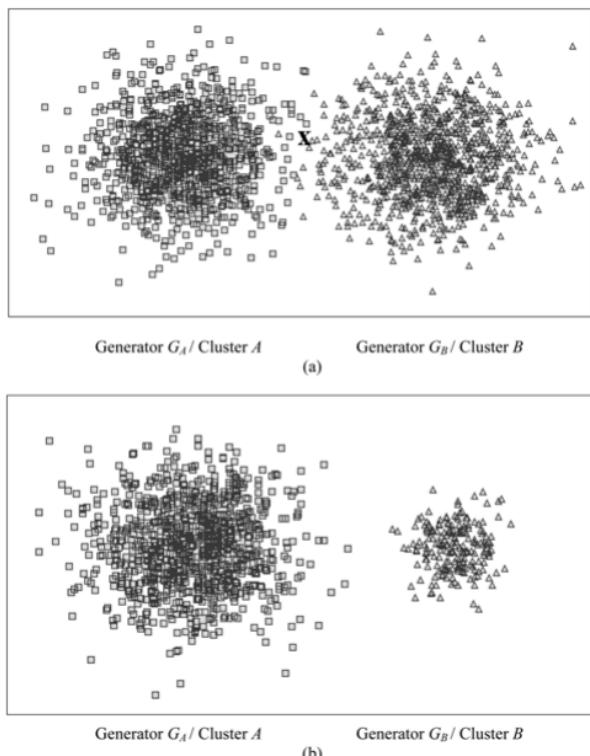


図1 クラスタ間の関係がクラスタリングに与える影響 (a) 2つの異なる生成者によって生成されたオブジェクトを含む領域が存在し、(b) 重複する領域が存在しない2つのオブジェクト空間: A,  $G_A$ によって生成されたオブジェクト ; D,  $G_B$ によって生成されたオブジェクト。

したがって、確率的手法で使用される統計的尺度は、K-meansアルゴリズムでは適用できません。さらに、基礎となる分布に関する仮定は実際のデータセットでは検証できないため、統計的尺度を得るために使用することはできません。

## 2.6 クラス数と等しいKの値

この方法では、クラスタ数はデータセットのクラス数と等しい。クラス属性が省略されたデータセットにデータクラスタリングアルゴリズムを適用し、省略されたクラス情報を使用してクラスタリング結果を評価することで、データクラスタリングアルゴリズムを分類器として使用することができる[26, 27]。評価の結果はクラスタリング・アルゴリズムにフィードバックされ、そのパフォーマンスを向上させます。このように、クラスタリングは教師付きであると考えることができる。

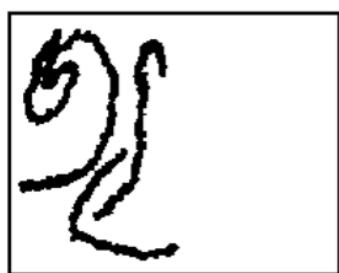
クラスタ数を決定するこの方法では、データクラスタリング手法がクラスタを形成でき、その各クラスタは1つのクラスに属するオブジェクトのみから構成されるとい

う仮定がなされる。残念ながら、ほとんどの現実の問題はこの仮定を満たさない。

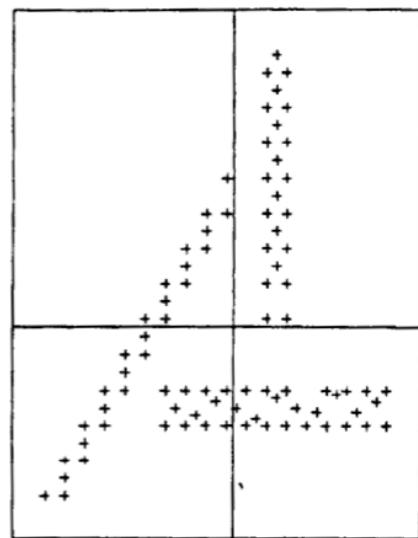
## 2.7 可視化によって決定されるKの値

視覚的検証は、その単純さと説明の可能性から広く適用されている。視覚的な例は、アルゴリズムの欠点を説明したり、期待されるクラスタリング結果を提示するためによく使用される[5, 27]。

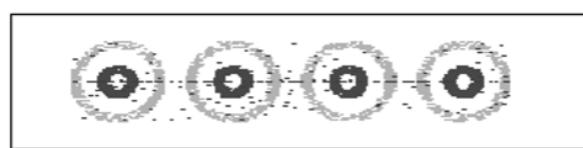
可視化技術を使用したクラスタリング結果の評価は、その暗黙的な性質に大きく依存する。クラスタリング手法によって利用されるクラスタリングモデルは、特定のデータセットに適切でない場合がある。図2のデータセットはそのようなケースの例である。可視化手法の適用は、予想されるクラスタにおけるデータ分布の連続性を意味する。このようなデータセットにK-meansアプローチを適用した場合、



(a)



(b)



(c)

図2 K平均法に不適切なデータセット：(a)4つのクラスターを持つデータセット[5]、(b)3つのクラスターを持つデータセット[23]、(c)8つのクラスターを持つデータセット[27]。各データセットのクラスタ数は、それぞれの研究者が指定したものである。



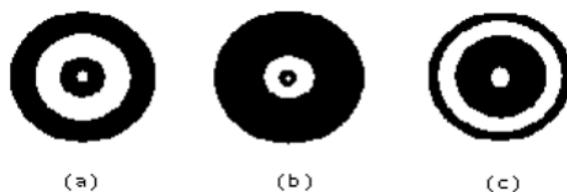


図3 2つのリングのデータセットにおける変化

K-meansクラスタリング・モデルを満たし、同時に例示されたデータセット内の特定のオブジェクト・グループに対応するクラスタは存在しない。したがって、K平均アルゴリズムは期待されるクラスタリング結果を得ることができない。このことは、K-meansアプローチはこのようなデータセットには適さないことを示唆している。

図2のデータセットの特徴（位置、形状、サイズ、オブジェクトの分布）は暗黙的に定義されている。これはクラスタリング結果の検証を困難にしている。データ特性にわずかな変化があれば、異なる結果につながる可能性がある。図2bのデータセットはそのようなケースの例である。もう一つの例は、図3の一連のデータセットである。図3aのデータセットでは2つのクラスターが容易に識別できるが、図3bとcのデータセットではクラスターの数がリング間の距離と各リングの物体密度に依存する。通常、このようなパラメータは、目視で確認する場合には明示的に定義されない。

上記の欠陥にもかかわらず、結果の視覚化は、データセットがクラスタリングモデルの仮定に違反していない場合、Kを選択しクラスタリング結果を検証するための有用な方法である。さらに、この方法は、期待される結果が明示的に特定できる場合に推奨される。

## 2.8 近傍尺度を使用して決定されたKの値

Kを決定するために、K-meansアルゴリズムのコスト関数に近傍尺度を追加することができる[26]。このテクニックはいくつかのデータセットで有望な結果を示したが、実用的なアプリケーションでその可能性を証明する必要がある。コスト関数を修正する必要があるため、このテクニックはオリジナルのK-meansアルゴリズムには適用できない。

$k$  の選択に影響を与える3つの要因

クラスタリング結果を評価する関数 $f(K)$ は、クラスタ数を選択するために使用できる。そのような関数が考慮すべき要素については、この節で議論する。

### 3.1 Approach bias

評価関数はクラスタリング基準と密接に関連する必要があります。前述したように、このような関係があれば、検証プロセスへの悪影響を防ぐことができる。特にK-meansアルゴリズムでは、クラスタの歪みを最小化することが基準であるため、評価関数はこのパラメータを考慮する必要がある。

### 3.2 Level of detail

一般的に、比較的低い詳細度しか見えない観察者は、物体の概観しか得られない。詳細度を上げることで、観察対象物に関するより多くの情報を得ることができるが、同時に処理しなければならないデータ量も増える。リソースの制約があるため、通常、高い詳細レベルは対象物の一部を検査するためにのみ使用される[28]。

このようなアプローチはクラスタリングに適用できる。n個のオブジェクトを持つデータセットは、1からnの間の任意の数のクラスターにグループ化することができる。異なるK値を指定することで、オブジェクトをさまざまな数のクラスターにグループ化した結果を評価することができる。この評価から、複数のK値がユーザーに推奨される可能性があるが、最終的な選択はユーザーが行う。

### 3.3 内部分布対グローバルインパクト

クラスタリングは、データ分布の不規則性を見つけ、オブジェクトが集中している領域を識別するために使用される。しかし、オブジェクトが集中しているすべての領域がクラスターとみなされるわけではない。ある領域がクラスターとして識別されるためには、その内部分布だけでなく、データセット内の他のオブジェクトのグループ化との相互依存性も分析することが重要である。

K-meansクラスタリングでは、クラスタの歪みはデータ母集団とオブジェクトとクラスタ中心間の距離の関数である。

$$I_j = \sum_{t=1}^{N_j} [d(x_{jt}, w_j)]^2 \quad (1a)$$

ここで、 $I_j$  はクラスタ  $j$  の歪み、 $w_j$  はクラスタ  $j$  の中心、 $N_j$  はクラスタ  $j$  に属する物体の数、 $x_{jt}$  はクラスタ  $j$  に属する  $t$  番目の物体、 $d(x_{jt}, w_j)$  は物体  $x_{jt}$  とクラスタ  $j$  の中心  $w_j$  との距離である。この式は、各クラスタの歪みを表現するための指標である。

次式で与えられる全歪みの合計  $S_K$  に対する寄与によって評価されます。

$$S_K = \sum_{j=1}^K I_j \quad (1b)$$

ここでは  $K$  は指定されたクラスタ数である。

したがって、このような情報は、オブジェクト空間の特定の領域がクラスターとみなされるかどうかを評価する上で重要である。

### 3.4 $f(K)$ の制約条件

$f(K)$  のロバスト性は非常に重要である。この関数はクラスタリングアルゴリズムの結果に基づいているため、 $K$  が変化しない場合、この結果ができるだけ変化しないことが重要である。しかし、K-means アプローチの主な欠陥の1つは、ランダム性に依存していることです。したがって、その性能を評価関数の変数として使用できるように、アルゴリズムは一貫した結果をもたらすべきである。Kmeansアルゴリズムの新しいバージョン、すなわちインクリメンタルK-meansアルゴリズム[29]は、この要件を満たし、この目的に採用できる。

$f(K)$  の役割は、データ分布の傾向を明らかにすることであり、したがって、オブジェクトの数に依存しないことが重要である。クラスタ数  $K$  はオブジェクト数  $N$  よりもはるかに小さいと仮定します。 $K$  が増加するとき、 $f(K)$  はある一定の値に収束するはずです。そして、任意の中間  $K$  に対して、 $f(K)$  が最小点や最大点のような特別な振る舞いを示す場合、その  $K$  の値を望ましいクラスタ数とみなすことができる。

## 4 NUMBER OF CLUSTERS FOR K-MEANS CLUSTERING

3.3節で述べたように、クラスター分析はデータ分布の不規則性を見つけるために使用される。データ分布が一様であれば、不規則性はない。したがって、一様な分布を持つデータセットは、クラスタリング結果の校正と検証に使用することができる。このアプローチは Tibshirani らによって適用された[30]。実際のデータセットと同じ次元で、一様な分布を持つデータセットが生成された。そして、この人工データ集合のクラスタリング性能を、実際のデータ集合で得られた結果と比較した。性能を評価するために、「ギャップ」統計量[30]として知られる尺度が採用された。この作業では、人工データ集合を生成する代わりに、人工データ集合に対するクラスタリング性能を推定した。

また、ギャップ統計量の代わりに、クラスタリング結果を評価するためにより識別性の高い新しい尺度が採用された。

K-means アルゴリズムが一様分布のデータに適用され、 $K$  が1増加すると、クラスタが変化し、新しい位置では、パーティションは再びほぼ等しいサイズとなり、それらの歪みは互いに類似する。参考文献[29]で行われた評価では、超立方体形状で一様分布を持つクラスタ ( $K \leq 1$ ) に新しいクラスタを挿入した場合、歪みの総和の減少は元の歪みの総和に比例することが示されました。この結論は、比較的小さな  $K$  の値で得られたクラスタリング結果に対して正しいことがわかった。このような場合、クラスタ数を増やした後の歪みの総和は、現在の値から推定することができる。

評価関数  $f(K)$  は次式で定義される。

$$f(K) = \begin{cases} 1 & \text{if } K = 1 \\ \frac{S_K}{\alpha_K S_{K-1}} & \text{if } S_{K-1} \neq 0, \forall K > 1 \\ 1 & \text{if } S_{K-1} = 0, \forall K > 1 \end{cases} \quad (2)$$

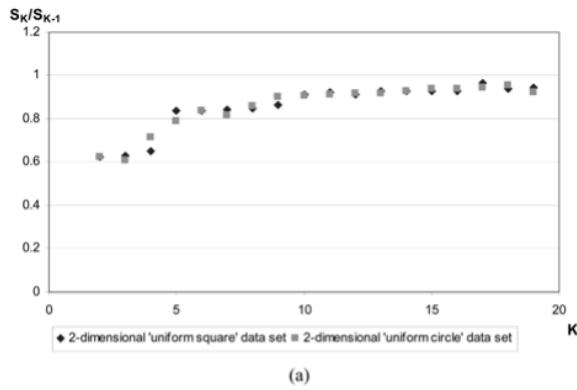
$$\alpha_K = \begin{cases} 1 - \frac{3}{4N_d} & \text{if } K = 2 \text{ and } N_d > 1 \end{cases} \quad (3a)$$

$$\alpha_K = \alpha_{K-1} + \frac{1 - \alpha_{K-1}}{6} \quad \text{if } K > 2 \text{ and } N_d > 1 \quad (3b)$$

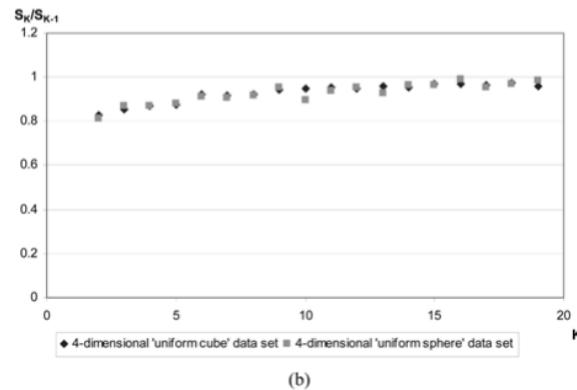
ここで  $S_K$  はクラスタ数が  $K$  の時のクラスタ歪みの総和、 $N_d$  はデータ集合の属性数（すなわち次元数）、 $\alpha_K$  は重み因子である。式(2)の  $\alpha_K S_{K-1}$  の項は、データが一様分布であるという仮定で作られた  $S_{K-1}$  に基づく  $S_K$  の推定値である。 $f(K)$  の値は、推定された歪みに対する実際の歪みの比であり、データ分布が一様であれば 1 に近くなる。データ分布に集中する部分があると、 $S_K$  は推定値より小さくなり、 $f(K)$  は小さくなります。 $f(K)$  が小さいほど、データ分布が集中していることになります。したがって、 $f(K)$  が小さくなる  $K$  の値は、よく定義されたクラスターを与えるとみなすことができます。

式(3)で定義される重み係数  $\alpha_K$  は、1以下の正数であり、次元の影響を小さくする ために適用される。 $K \leq 2$  のとき、 $\alpha_K$  は式(3a)を用いて計算される。この式は参考文献[29]の式(7)から導かれ、歪みの減少が次元数  $N_d$  に反比例することを示している。

$K$  が 2 より大きくなると、図4からわかるように、歪みの総和の減少が小さくなる（比  $S_K / S_{K-1}$  が 1 に近づく）。



(a)

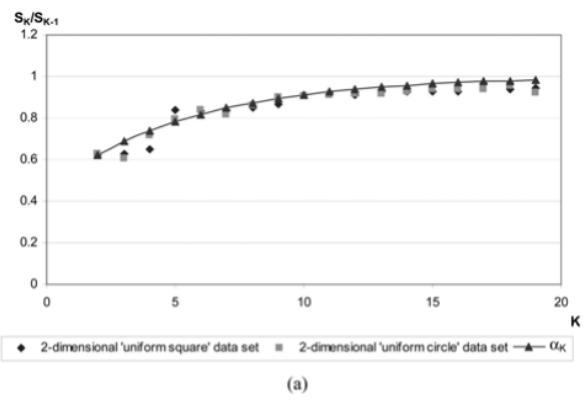


(b)

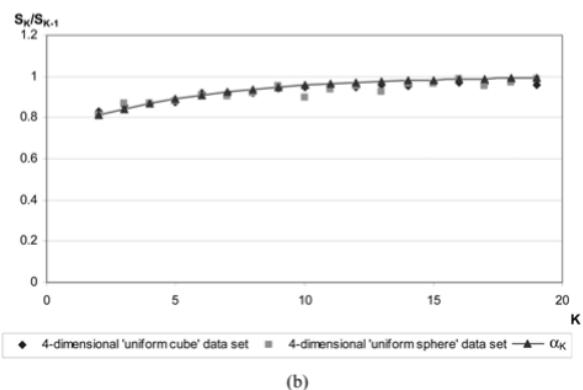
図4 一様な分布を持つデータセットに対する比率 $S_K / S_{K-1}$  (a) 2次元の「正方形」と「円」 (b)

この図は、異なる次元の一様分布のデータセットにクラスタリングアルゴリズムを適用したとき、異なるKについて計算された $S_K / S_{K-1}$ の値を示している。このようなデータセットでは、 $f(K)$ は1に等しいことが期待され、 $a_K$ は $f(K)$ が1に等しくなるように選択されるべきである。式(2)から $a_K$ は $S_K / S_{K-1}$ となり、図4から得られるはずである。しかし、計算を簡単にするために、図4のデータから再帰式(3b)を導出し、 $a_K$ を計算している。図5から、式(3b)から得られる $a_K$ の値は、図4のプロットに密接にフィットすることがわかる。提案関数 $f(K)$ は前節で述べた制約を満たす。 $f(K)$ のロバスト性は次節で実験的に検証する。オブジェクトの数が2倍、3倍になっても、その分布が変わらない場合、結果として得られるクラスタは同じ位置に留まる。 $S_K$ と $S_{K-1}$ はそれに応じて2倍または3倍になるので、 $f(K)$ は一定である。したがって、一般的に $f(K)$ はデータセットのオブジェクト数に依存しない。

属性の範囲の違いの影響を減らすために、データはクラスタリングを開始する前に正規化される。しかし、注意すべき点がある、



(a)



(b)

図5 式(3b)を用いて計算した $a_K$ の値と比率 $S_K / S_{K-1}$ の比較

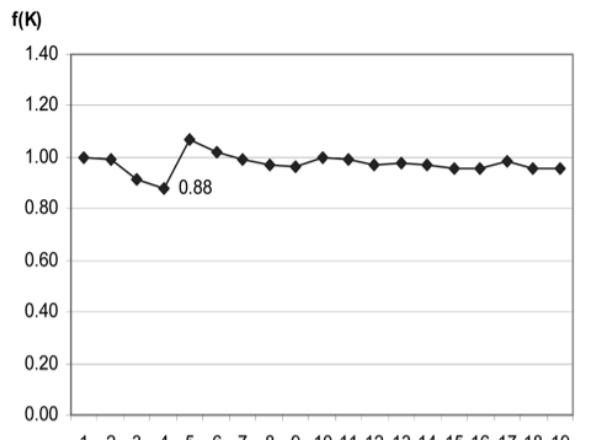
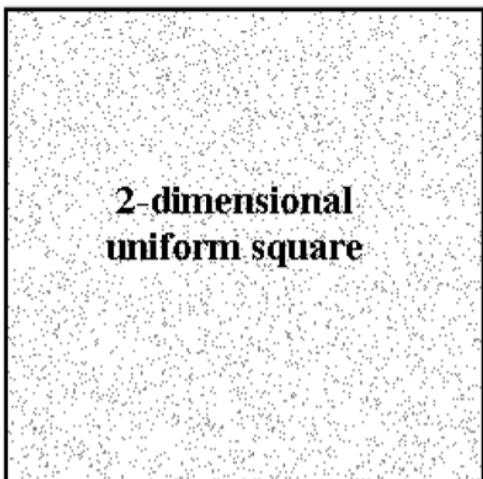
4次元の「立方体」と「球体」オブジェクトのグループがよく分離されたデータでは、問題空間における そのような領域の形状が評価関数に影響を与える。このような場合、正規化はデータセット全体に適用されるスケーリング技法であるため、局所的なオブジェクト分布に影響を与えない。

## 5 PERFORMANCE

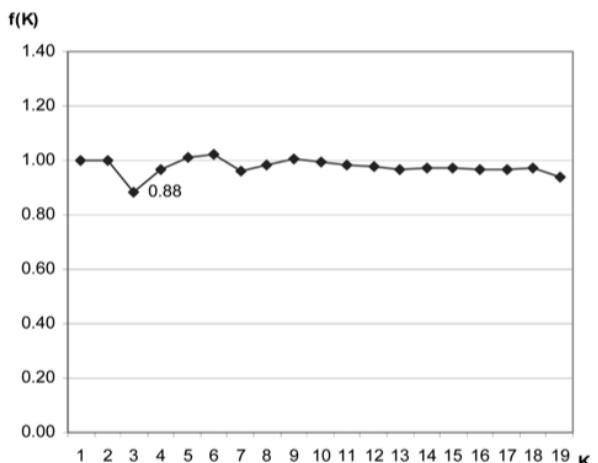
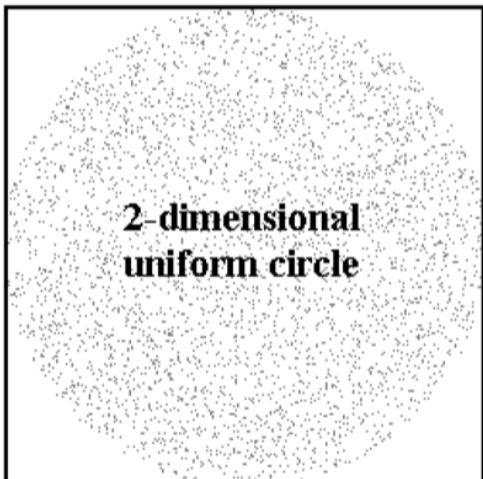
評価関数 $f(K)$ は、図6に示す人工的に生成されたデータセットで、一連の 実験でテストされる。 $f(K)$ はクラスタの歪みの合計に基づいて計算される。

図6a～cでは、すべてのオブジェクトは一様な分布を持つ単一の領域に属している。図6aのグラフは、 $f(K)$ がすべてのKに対してほぼ一定で1に等しいため、 $f(K)$ が一様分布のこのデータ集合のクラスタリング結果をよく反映していることを示している。図6aおよびbにおいて、それぞれK ≈ 4およびK ≈ 3のとき、 $f(K)$ は最小値に達する。これは、これらのデータセットに属する物体によって定義される領域の形状に起因すると考えられる。しかし、 $f(K)$ の最小値に大きな違いはない。

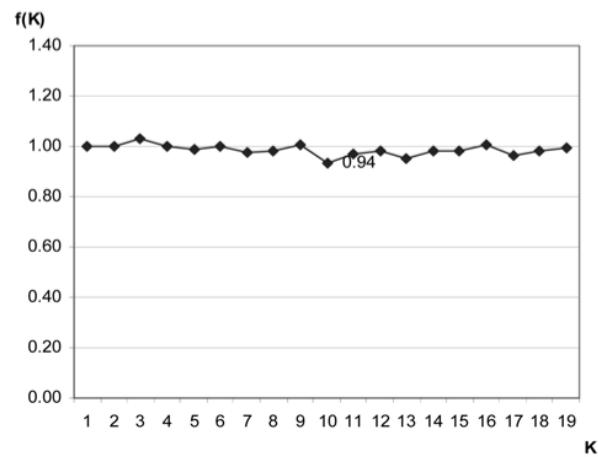
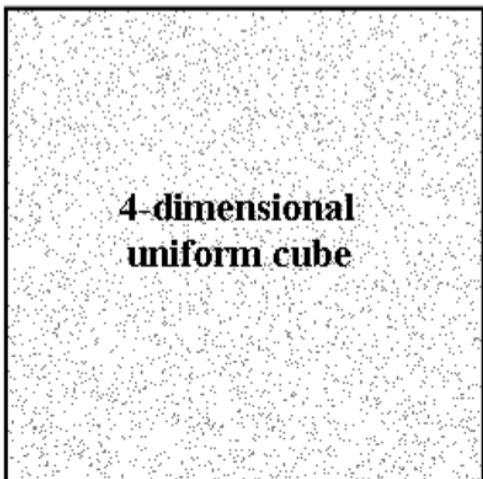




(a)

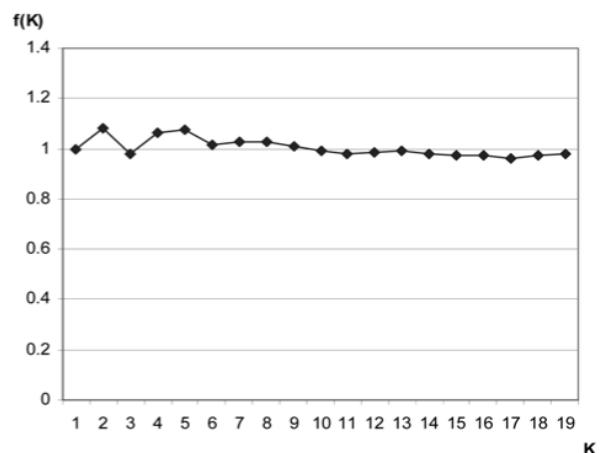
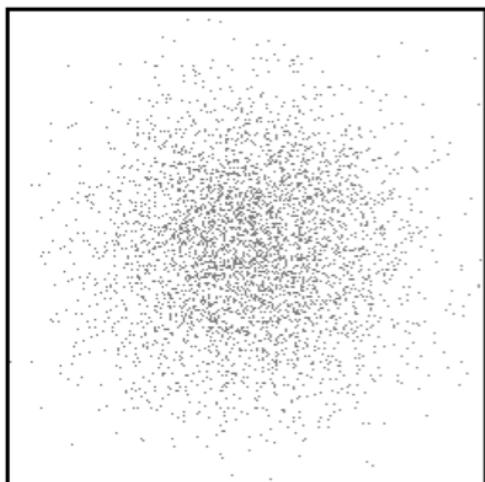


(b)

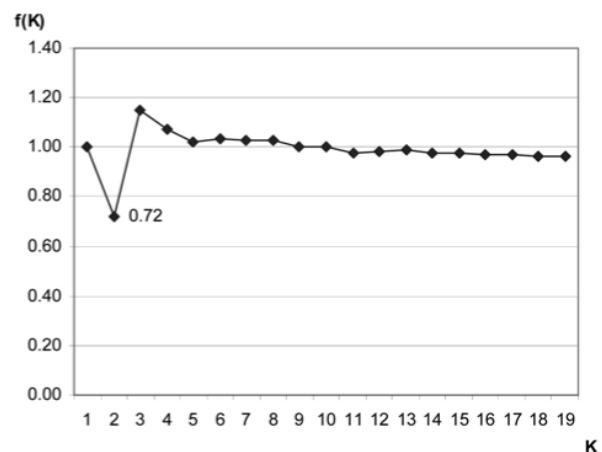
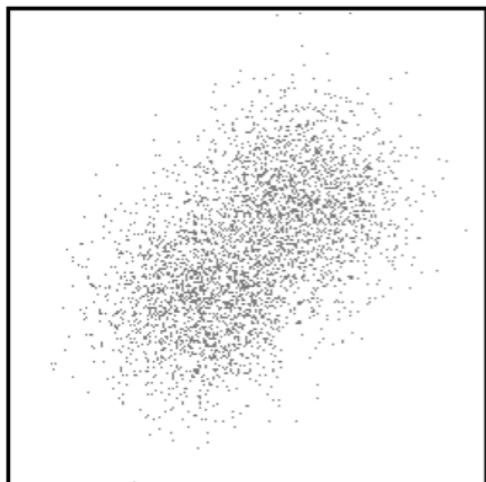


(c)

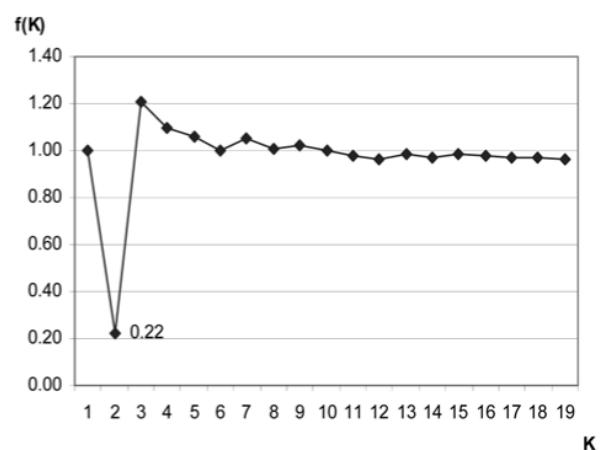
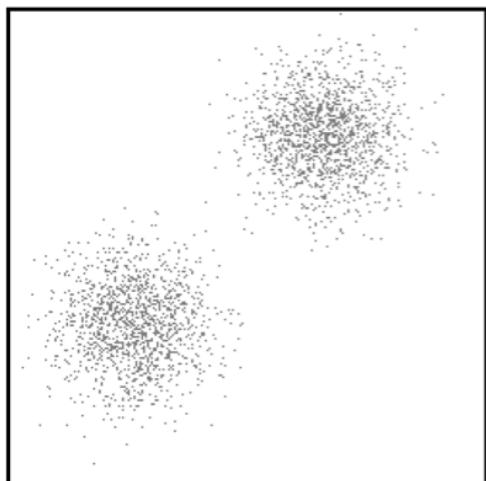
**Fig. 6** Data sets and their corresponding  $f(K)$



(d)

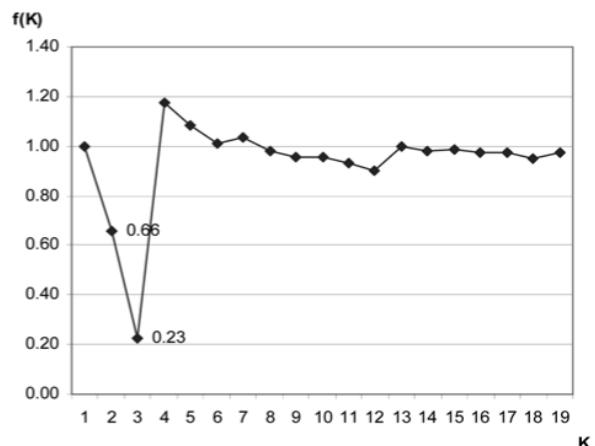
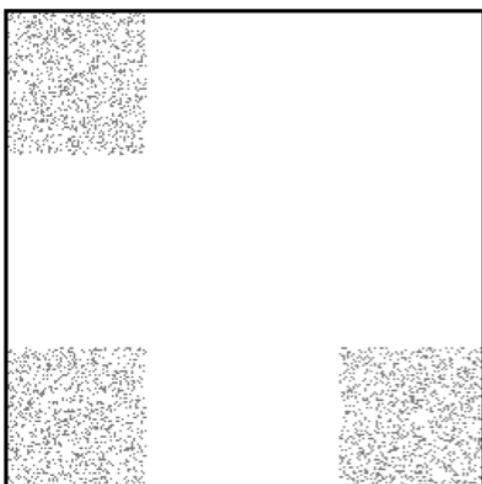


(e)

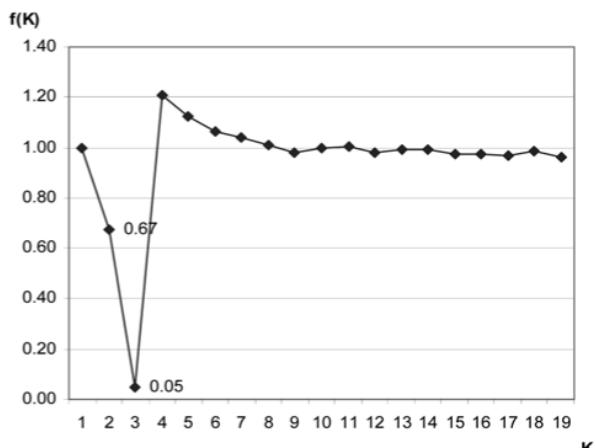
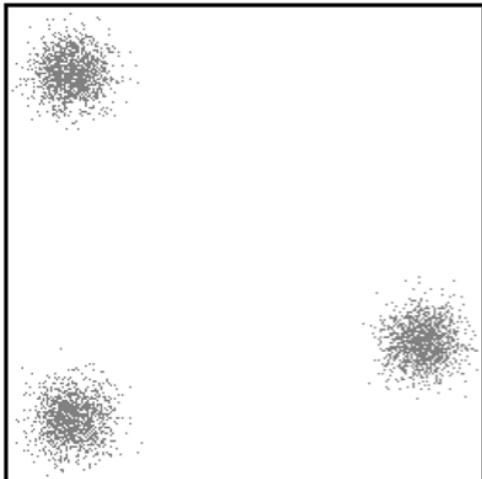


(f)

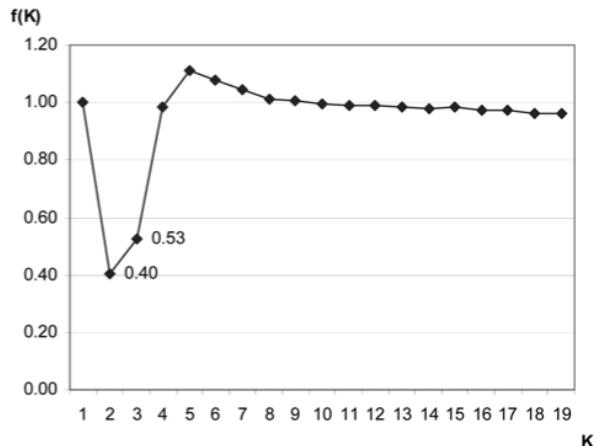
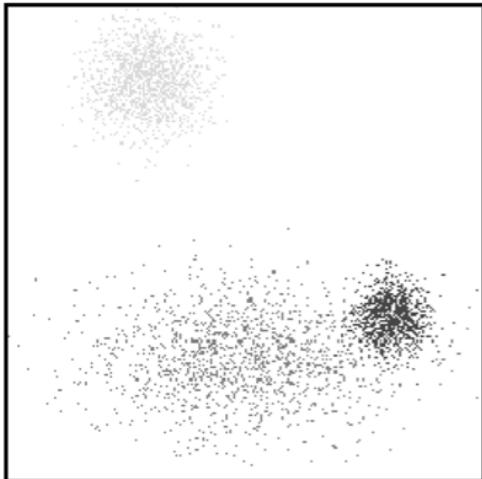
Fig. 6 Continued



(g)



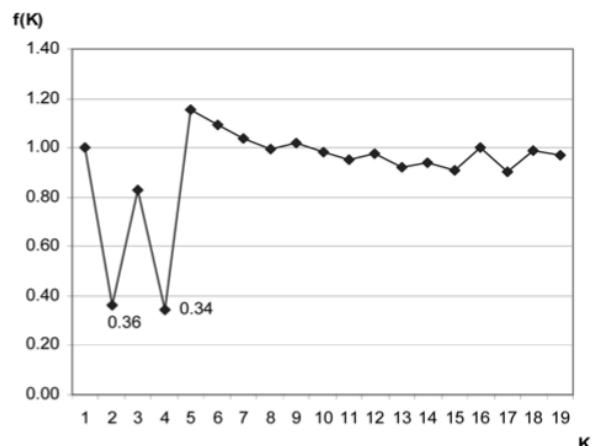
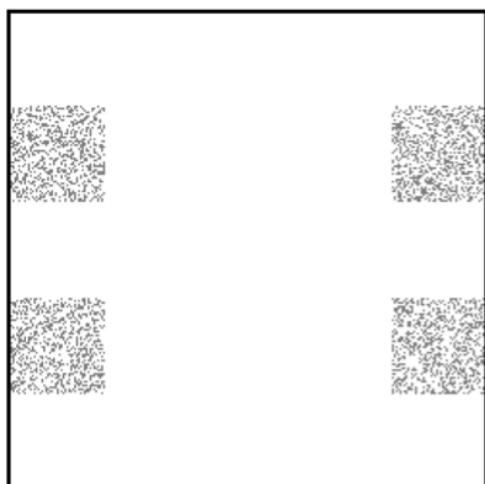
(h)



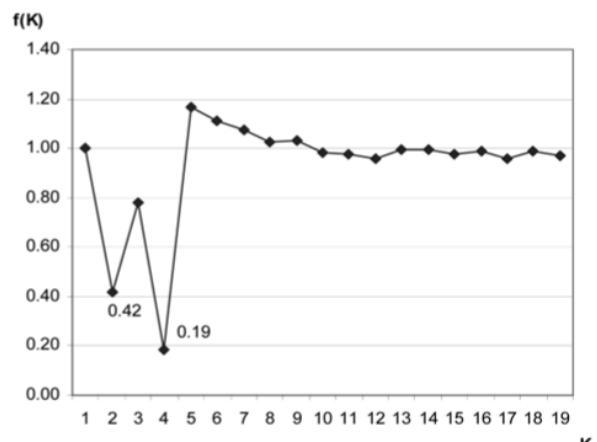
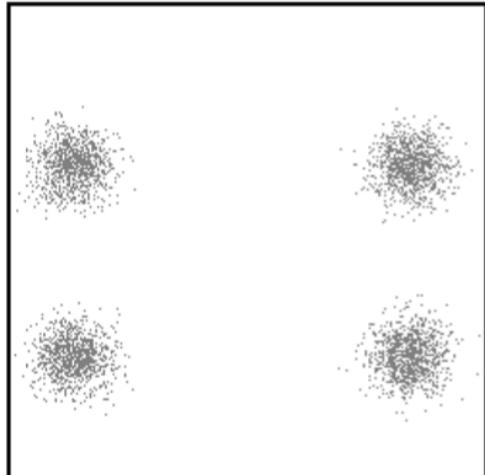
(i)

Fig. 6 Continued

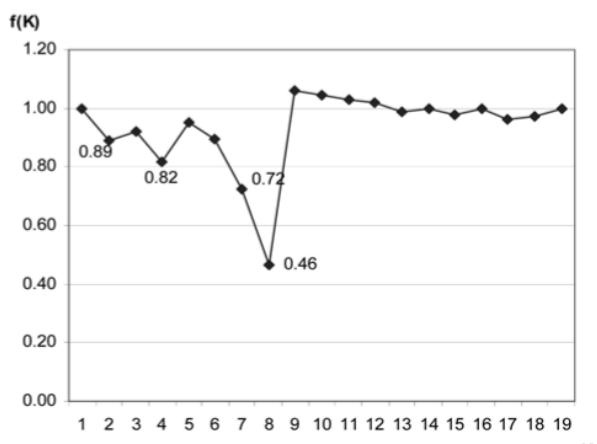
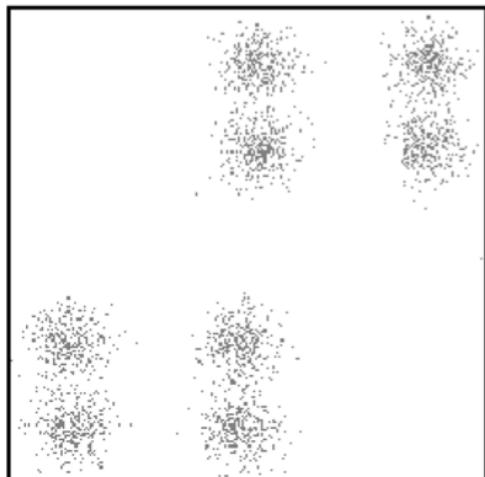




(j)



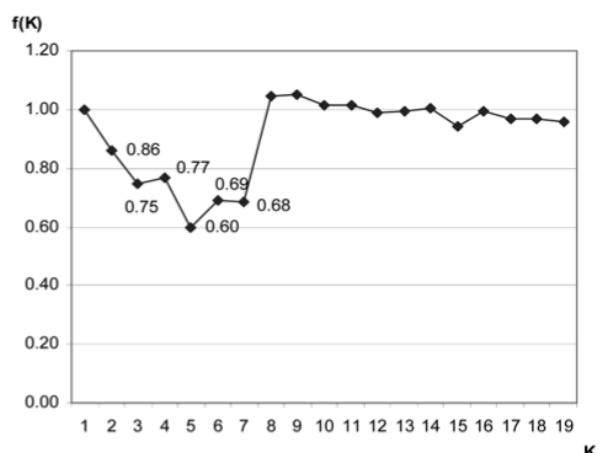
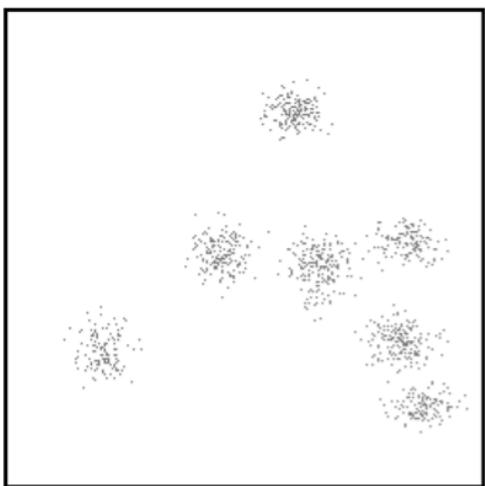
(k)



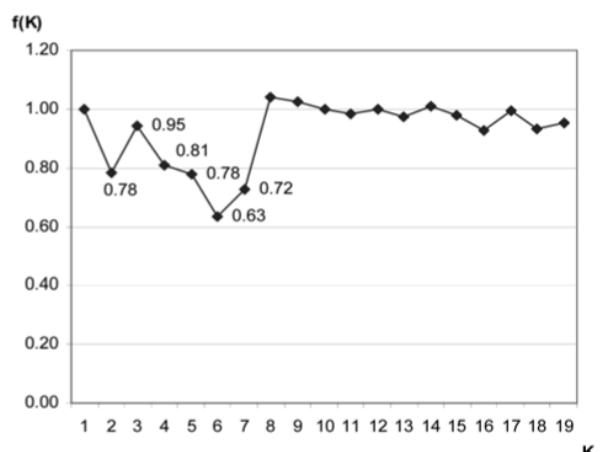
(l)

Fig. 6 Continued

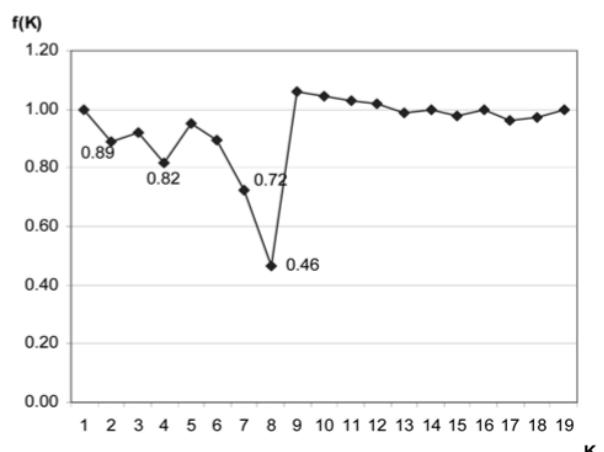
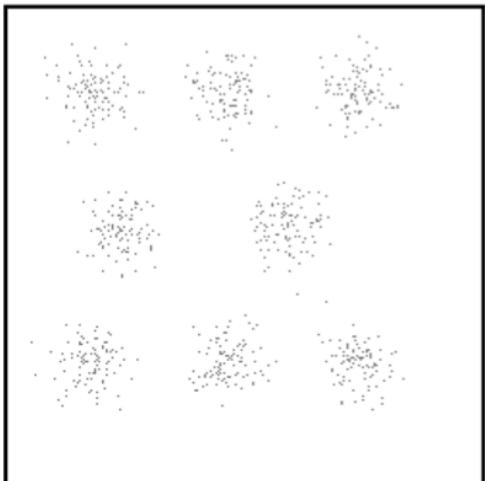




(m)



(n)



(o)

Fig. 6 Continued

ユーザーへの強力な推奨を行うために、平均値から $\#$ を選択する。図6aとcの $f(K)$ の値を比較することで、 $a_K$ はデータセットの次元が評価関数に与える影響を軽減することがわかる。

図6dのデータセットでは、ここでもまた、すべてのオブジェクトが、正規分布を持つ単一の領域に集中している。

このデータセットの $f(K)$ プロットは、 $K \leq 1$ のとき、クラスタリング結果がこのデータセットに最も適していることを正しく示唆している。

図6eとfのデータセットは、正規分布を持つ2つのジェネレーターによって作成されている。図6eでは、2つのジェネレーターは重なり合う領域を持っているが、図6fでは、

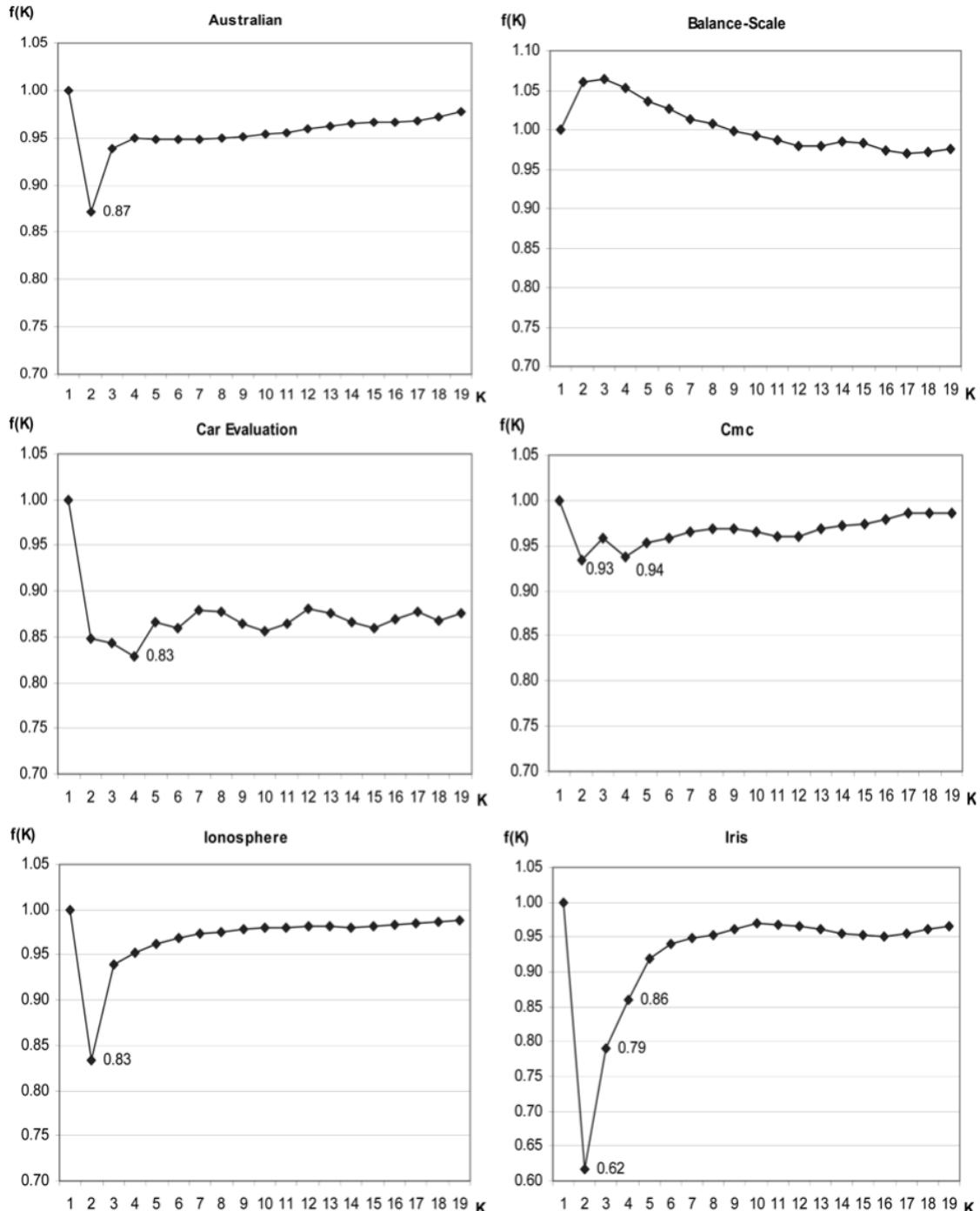


図7 12のベンチマークデータセットに対する $f(K)$

よく分離されている。後者の図の $f(2)$ の値は、前者よりもずっと小さいことに注意。

図6gとhのデータセットには、3つの認識可能な領域がある。対応するグラフから、 $f(K)$ はこれらのデータセットをクラスタリングするための $K$ の正しい値を示唆している。

図6iのデータセットを形成するために、正規分布を持つオブジェクトグループを作成する3つの異なるジェネレータが使用されている。この場合、 $f(K)$ は $K$ に対して2または3の値を示唆する。これら3つのジェネレーターのうち2つはオブジェクトのグループ化を作成し重複するため、 $f(2)$ は $f(3)$ よりも小さい。

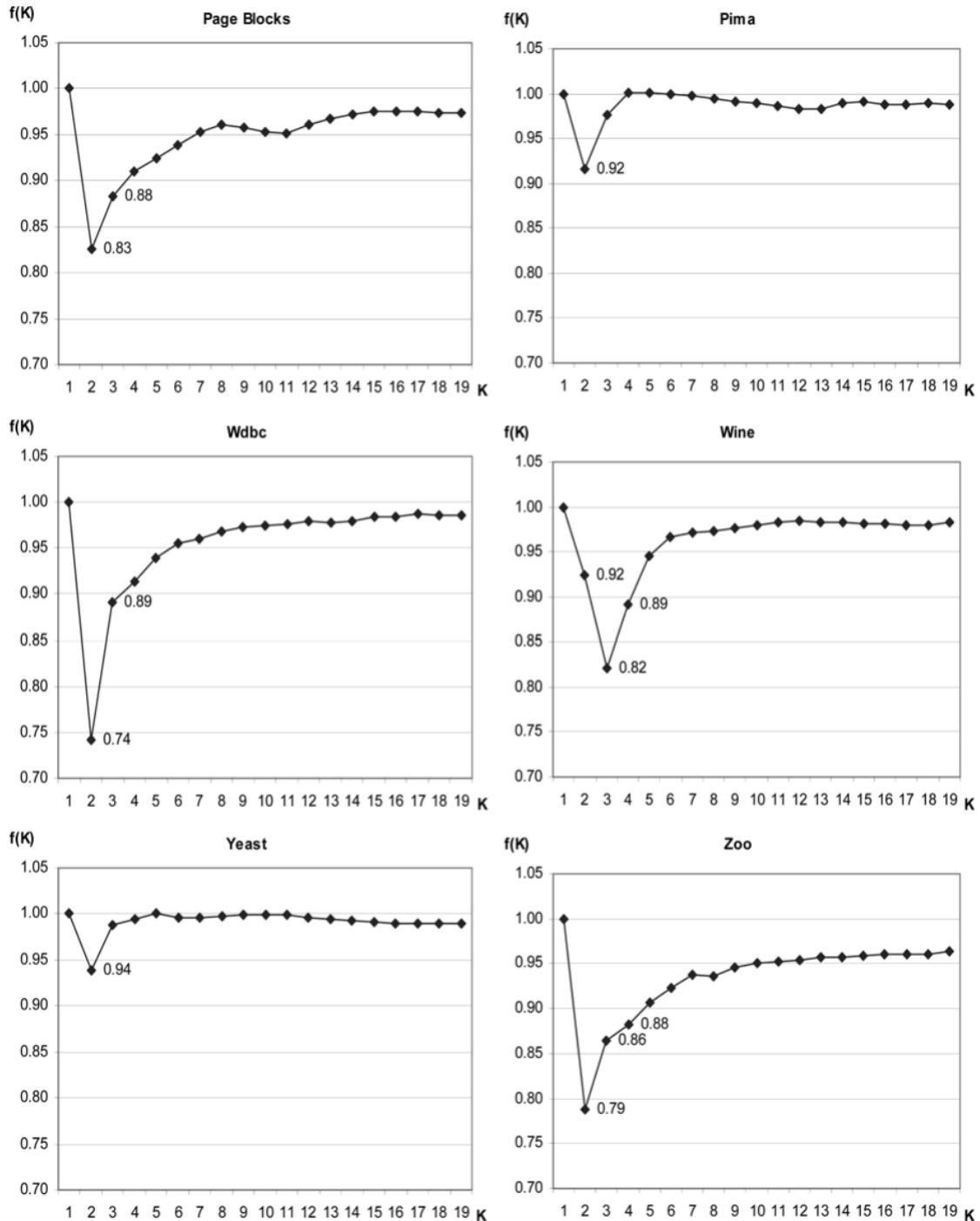


Fig. 7 Continued

これは、データには明確に定義された領域が2つしかないことを意味するが、 $K \leq 3$ もオブジェクトのクラスタリングに使用できる。 $f(K)$ はそれぞれ  $K \leq 2$  と 4 で最小値に達する。このような場合、ユーザーは特定の要件に基づき、最も適切なKの値を選択することができます。より複雑なケースを図6lに示すが、Kの値は4または8の可能性がある。特定のKの選択は、クラスタリングが実行される特定のアプリケーションの要件に依存する。

図6m~oのデータセットは、オブジェクト空間においてよく定義された領域を持っており、それぞれの領域は、オブジェクトの分布、位置、数が異なっている。オブジェクトのクラスタリングに  $f(K)$  の最小値を使用する場合、Kは、それらを作成するために利用されたジェネレータの数（図6oのクラスタの場合のように）、または視覚的に識別できたオブジェクトのグループ化の数（図6mとnのクラスタの場合のように）とは異なる。この違いの理由はケースによって異なる。例えば、図6mでは、左端の2組のクラスターの距離が他のクラスターよりも小さく、それらの組のクラスターを統合することができたので、5つのクラスターがあると考えることができる。このことは、 $f(K)$  はクラスター数の目安を示唆するために使われるべきであり、どの値を採用するかの最終的な決定はユーザーの裁量に委ねられるべきであるという事実を浮き彫りにしている。

図6のグラフから、対応する  $f(K)$  , 0.85を持つ任意のKがクラスタリングに推奨できるという結論が得られる。対応する  $f(K)$  , 0.85を持つ値がない場合、 $K \leq 1$  が選択される。

提案された関数  $f(K)$  は、UCI Repository Machine Learning Databases [31]の12個のベンチマークデータセットにも適用された。図7は、 $f(K)$  の値がKによってどのように変化するかを示している。 $f(K)$  に0.85の閾値が選択された場合（人工データセットの研究から）、これらのデータセットのそれぞれに対して推奨されるクラスター数は表2のように与えられる。 $K \leq 1$  は、データ分布が標準一様分布に非常に近いことを意味する。 $f(K)$  を使って推奨された値は、これらのデータセットの属性間の相関が高いため、非常に小さく、図6eに示されたものと非常によく似ている。これは、一度に2つの属性を調べ、データセットを2次元にプロットすることで検証できる。

15個の人工データセットと12個のベンチマークデータセットを用いた上記の実験的研究により、 $f(K)$  の頑健性が実証された。評価関数は、Kが9より大きくなると、ほとんどの場合1に収束する。

表2  $f(K)$ に基づく推奨クラスター数

Data sets	推奨クラスター数
Australian	1
Balance-scale	1
Car evaluation	2, 3, 4
Cmc	1
Ionosphere	2
Iris	2, 3
Page blocks	2
Pima	1
Wdbc	2
Wine	3
Yeast	1
Zoo	2

## 6 CONCLUSION

K平均クラスタリングのクラスター数を選択する既存の方法には、多くの欠点があります。また、クラスタリング結果を評価するための現在の方法は、クラスタリングアルゴリズムの性能に関する多くの情報を提供しない。

この論文では、K-means アルゴリズムのクラスター数を選択する新しい方法が提案されている。この新しい方法は、アルゴリズムの性能を反映する情報を考慮するため、K-meansクラスタリングのアプローチと密接に関連している。提案手法は、様々な要求される詳細度で異なるクラスタリング結果が得られるような場合に、ユーザーに複数のK値を提案することができる。この方法は、Kのガイド値を提案する前にK-meansアルゴリズムを数回適用する必要があるため、大規模なデータセットで使用すると計算コストが高くなる可能性がある。より複雑なオブジェクト分布を持つデータセットにこの方法を適用した場合の能力を検証するために、さらなる研究が必要である。

## ACKNOWLEDGEMENTS

この研究は、工学・物理科学研究評議会の支援を受けたカーディフ・イノベティブ・マニュファクチャリング研究センター・プロジェクト、および欧州地域開発基金プログラムの下で欧州委員会とウェールズ議会政府の支援を受けたSUPERMANプロジェクトの一環として実施された。著者らは欧州委員会の資金援助を受けたI PROMS Network of Excellenceのメンバーである。

## REFERENCES

- 1** Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*, 2000 (Morgan Kaufmann, San Francisco, California).
- 2** Al-Daoud, M. B., Venkateswarlu, N. B., and Roberts, S. A. Fast  $K$ -means clustering algorithms. Report 95.18, School of Computer Studies, University of Leeds, June 1995.
- 3** Al-Daoud, M. B., Venkateswarlu, N. B., and Roberts, S. A. New methods for the initialisation of clusters. *Pattern Recognition Lett.*, 1996, **17**, 451–455.
- 4** Alsabti, K., Ranka, S., and Singh, V. An efficient  $K$ -means clustering algorithm. In Proceedings of the First Workshop on High-Performance Data Mining, Orlando, Florida, 1998; <ftp://ftp.cise.ufl.edu/pub/faculty/ranka/Proceedings>.
- 5** Bilmes, J., Vahdat, A., Hsu, W., and Im, E. J. クラスタリング問題に対する確率的ヒューリスティックの経験的観測。技術報告TR-97-018, 国際計算機科学研究所, カリフォルニア州パークレー。
- 6** Bottou, L. and Bengio, Y. Convergence properties of the  $K$ -means algorithm. *Adv. Neural Infn Processing Systems*, 1995, **7**, 585–592.
- 7** Bradley, S. and Fayyad, U. M. Refining initial points for  $K$ -means clustering. このような場合、 $K$ -Meansクラスタリングでは、 $K$ -Meansクラスタリングの初期点を選択する必要がある。
- 8** Du, Q. and Wong, T-W. MacQueen's  $K$ -means algorithm for computing the centroidal Voronoi tessellations. *Int. J. Computers Math. Applications*, 2002, **44**, 511–523.
- 9** (注1)本論文は、「データ工学と自動学習」(IDEAL 2000), 香港, 中国, 2000年12月, pp.17–22. 第4回データベースとデータマイニングにおける知識発見の原理に関するヨーロッパワークショップ(PKDD 00), リヨン, フランス, 2000, pp. 17–22.
- 10** Castro, V. E. Why so many clustering algorithms? *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2002, **4**(1), 65–75.
- 11** Fritzke, B. The LBG-U method for vector quantization – an improvement over LBG inspired from neural networks. *Neural Processing Lett.*, 1997, **5**(1), 35–45.
- 12** Hamerly, G. and Elkan, C. Alternatives to the  $K$ -means algorithm that find better clusterings. In Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 02), McLean, Virginia, 2002, pp. 600–607.
- 13** Hansen, L. K. and Larsen, J. Unsupervised learning and generalisation. In Proceedings of the IEEE International Conference on Neural Networks, Washington, DC, June 1996, pp. 25–30 (IEEE, New York).
- 14** Ishioka, T. Extended  $K$ -means with an efficient estimation of the number of clusters. In Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2000), Hong Kong, China, December 2000, pp. 17–22.
- 15** Kanungo, T., Mount, D. M., Netanyahu, N., Piatko, C., Silverman, R., and Wu, A. The efficient  $K$ -means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Analysis Mach. Intell.* 2002, **24**(7), 881–892.
- 16** Pelleg, D. and Moore, A. Accelerating exact  $K$ -means algorithms with geometric reasoning. In Proceedings of the Conference on Knowledge Discovery in Databases (KDD 99), San Diego, California, 1999, pp. 277–281.
- 17** Pelleg, D. and Moore, A. X-means: extending  $K$ -means with efficient estimation of the number of clusters. In Proceedings of the 17th International Conference on Machine Learning (ICML 2000), Stanford, California, 2000, 727–734.
- 18** Pena, J. M., Lazcano, J. A., and Larrañaga, P. An empirical comparison of four initialisation methods for the  $K$ -means algorithm. *Pattern Recognition Lett.*, 1999, **20**, 1027–1040.
- 19** SPSS Clementine Data Mining System. User Guide Version 5, 1998 (Integral Solutions Limited, Basingstoke, Hampshire).
- 20** DataEngine 3.0 – Intelligent Data Analysis – an Easy Job, Management Intelligenter Technologien GmbH, Germany, 1998; <http://www.mitgmbh.de>.
- 21** Kerr, A., Hall, H. K., and Kozub, S. *Doing Statistics with SPSS*, 2002 (Sage, London).
- 22** S-PLUS 6 for Windows Guide to Statistics, Vol. 2, Insightful Corporation, Seattle, Washington, 2001; <http://www.insightful.com/DocumentsLive/23/44/statman2.pdf>.
- 23** Hardy, A. On the number of clusters. *Comput. Statist. Data Analysis*, 1996, **23**, 83–96.
- 24** Theodoridis, S. and Koutroumbas, K. *Pattern Recognition*, 1998 (Academic Press, London).
- 25** Halkidi, M., Batistakis, Y., and Vazirgiannis, M. Cluster validity methods. Part I. *SIGMOD Record*, 2002, **31**(2); available online <http://www.acm.org/sigmod/record/>.
- 26** Kothari, R. and Pitts, D. On finding the number of clusters. *Pattern Recognition Lett.*, 1999, **20**, 405–416.
- 27** Cai, Z. Technical aspects of data mining. PhD thesis, Cardiff University, Cardiff, 2001.
- 28** Lindeberg, T. *Scale-space Theory in Computer Vision*, 1994 (Kluwer Academic, Boston, Massachusetts).
- 29** Pham, D. T., Dimov, S. S., and Nguyen, C. D. Incremental  $K$ -means algorithm. *Proc. Instn Mech. Engrs, Part C: J. Mechanical Engineering Science*, 2003, **218**, 783–795.
- 30** Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a dataset via the gap statistic. Technical Report 208, Department of Statistics, Stanford University, California, 2000.
- 31** Blake, C., Keogh, E., and Merz, C. J. UCI Repository of Machine Learning Databases, Irvine, California. Department of Information and Computer Science, University of California, Irvine, California, 1998.

## APPENDIX

## Notation

$A, B$	clusters	$N_j$	number of objects belonging to cluster $j$
$d(x_{jt}, w_j)$	distance between object $x_{jt}$ and the centre $w_j$ of cluster $j$	$P_{G_A}, P_{G_B}$	probabilities that $X$ is created by $G_A$ or $G_B$ respectively
$f(K)$	evaluation function	$P_{C_A}, P_{C_B}$	$X$ がそれAまたはBにクラスタ化される確率
$G_A, G_B$	generators	$S_K$	sum of all distortions with $K$ being the specified number of clusters
$I_j$	distortion of cluster $j$	$X$	object
$K$	number of clusters	$x_{jt}$	object belonging to cluster $j$
$N$	number of objects in the data set	$w_j$	centre of cluster $j$
$N_d$	number of data set attributes (the dimension of the data set)	$\alpha_K$	weight factor