(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2019年2月7日(07.02.2019)

WIPO | PCT

(10) 国際公開番号
WO 2019/026523 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)

(21) 国際出願番号： PCT/JP2018/025498

(22) 国際出願日： 2018年7月5日(05.07.2018)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2017-151146 2017年8月3日(03.08.2017) JP

(71) 出願人: 国立大学法人東北大学 (TOHOKU UNIVERSITY) [JP/JP]; 〒9808577 宮城県仙台市青葉区片平二丁目１番１号 Miyagi (JP).

(72) 発明者: 遠藤 哲郎 (ENDOH Tetsuo); 〒9808577 宮城県仙台市青葉区片平二丁目１番１号 国立大学法人東北大学内 Miyagi (JP). 沈 暉 (SHEN Hui); 〒9808577 宮城県仙台市青葉区片平二丁目１番１号 国立大学法人東北大学内 Miyagi (JP).

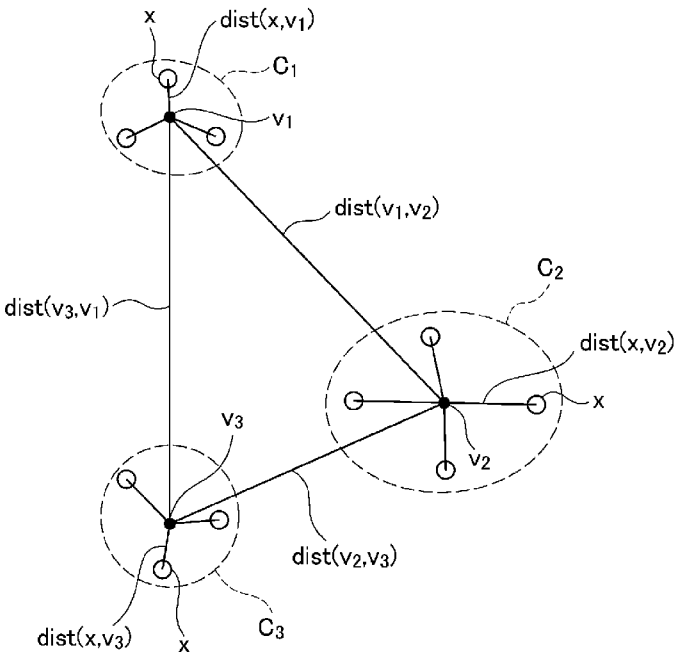
(74) 代理人: 吉田 正義 (YOSHIDA Tadanori); 〒1600023 東京都新宿区西新宿六丁目１５番１号 ラ・トゥール新宿４１１号 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

WO 2019/026523 A1

(54) **Title:** METHOD FOR CALCULATING CLUSTERING EVALUATION VALUE, AND METHOD FOR DETERMINING NUMBER OF CLUSTERS

(54) 発明の名称： クラスタリングの評価値算出方法及びクラスタ数決定方法



(57) **Abstract:** Provided are: a method for calculating a clustering evaluation value, the method making it possible to obtain an evaluation value for evaluating a clustering state with which it is possible to suppress an excessive number of classifications; and a method for determining the number of clusters, the method using the evaluation value. An evaluation value computation unit: derives, as a degree of internal connectedness, the sum for each cluster of the values obtained by normalizing the values indicating the degree of dispersion of the data within the cluster, using the number of items

[続葉有]

HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一 国際調査報告 (条約第21条(3))

of data in the cluster; derives a degree of external separation obtained by normalizing the value indicating the degree of mutual separation of the clusters, using the number of clusters; and obtains an evaluation value as a ratio of the degree of internal connectedness to the degree of external separation. The value indicating the degree of dispersion of the data within a cluster is the sum of squares of the distance within the cluster between the center of gravity of the cluster, and the data of the cluster. The degree of external separation is the sum of squares of the minimum distance between clusters derived for each cluster.

(57) 要約:過剰な分類を抑制することができるクラスタリング状態を評価する評価値を得ることができるクラスタリングの評価値算出方法及びその評価値を用いたクラスタ数決定方法を提供する。評価値演算部は、クラスタ内の各データの分散の程度を示す値を当該クラスタのデータ数で規格化した値の各クラスタについての総和を内的結合度として求めるとともに、クラスタの相互の分離の程度を示す値をクラスタ数で規格化した外的分離度を求め、内的結合度と外的分離度との比を評価値とする。クラスタ内の各データの分散の程度を示す値は、クラスタの重心と当該クラスタの各データとのクラスタ内距離の二乗和としている。外的分離度は、各クラスタについて求められる最小のクラスタ間距離の二乗和としている。

明　細　書

発明の名称 ：

クラスタリングの評価値算出方法及びクラスタ数決定方法

技術分野

[0001] 本発明は、クラスタリングの評価値算出方法及びクラスタ数決定方法に関する。

背景技術

[0002] 分類対象の複数のデータの集合を複数のクラスタ（データ群）に分類するクラスタリング（クラスタ分析）が知られており、近年、画像解析、データマイニング、ビッグデータの解析等で頻繁に利用されている。クラスタリングでは、類似するデータ同士を同じクラスタに分類し、クラスタ間ではデータができるだけ類似しないように分類することによって、その分類結果からデータの傾向や特徴の抽出等を行うものである。

[0003] 複数のデータをクラスタに分類するクラスタリング手法としては、種々の方法が知られており、代表的なクラスタリングの手法の１つとしてｋ－ｍｅａｎｓ法が知られている。ｋ－ｍｅａｎｓ法では、予めクラスタ数ｋが設定され、例えばＮ個の全データのうちから任意のｋ個のデータが選択され、それらをｋ個のクラスタの重心の初期値とする（手順１）。次に、各データのそれぞれについて、当該データとの距離が最も短い重心のクラスタに分類する（手順２）。各クラスタについて、クラスタ内のデータの平均を新たなクラスタの重心に設定する（手順３）。そして、各クラスタの重心が変化しなくなるまで、手順２と手順３とを繰り返し行う。

[0004] 上記のようなクラスタリング手法は、予め設定されたクラスタ数のクラスタにデータを分類するものであって、クラスタ数を予め設定する必要がある。通常、最適なクラスタ数を決定するために、クラスタ数を変化させてクラスタリングした各結果のそれぞれについての評価値を求め、評価値が極値、あるいは最大値ないし最小値のクラスタ数を最適なクラスタ数としている。

評価値にも種々のものが提案されており、代表的なものの１つとしてＶＲＣ（以下、評価値ＶＲＣという）が知られている（例えば、非特許文献１）。

[0005] 評価値ＶＲＣは、全データ数で規格化された内的結合度に対するクラスタ数で規格化した外的分離度の比である。評価値ＶＲＣにおける外的分離度の規格化前の値は、全データの重心とクラスタの重心との距離に当該クラスタ内のデータ数で重み付けした各クラスタのそれぞれの値の総和である。また、評価値ＶＲＣにおける内的結合度の規格化前の値は、クラスタの重心と当該クラスタ内の各データとの距離のクラスタごとの和を、各クラスタについて合計したものである。

先行技術文献

非特許文献

[0006] 非特許文献1：CALINSKI, Tadeusz; HARABASZ, Jerzy. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 1974, 3.1:1-27.

発明の概要

発明が解決しようとする課題

[0007] ところで、上記評価値ＶＲＣを用いた場合、全データ内において、他のデータに対して、範囲の大きなデータのまとまりがある場合やデータの分布密度が高いデータのまとまりがある場合、その大きな範囲のデータのまとまりや分布密度が高いデータのまとまりを過剰に分類して、クラスタ数が大きくなる傾向がある。すなわち、そのような大きな範囲のデータのまとまりや分布密度が高いデータのまとまりを１個のクラスタに分類するよりも、複数のクラスタに分類した場合に、評価値ＶＲＣが大きくなる傾向にある。

[0008] 本発明は、上記事情を鑑みてなされたものであり、過剰な分類を抑制することができるクラスタリング状態を評価する評価値を得ることができるクラスタリングの評価値算出方法及びその評価値を用いたクラスタ数決定方法を

提供することを目的とする。

課題を解決するための手段

[0009] 本発明は、複数のデータをクラスタリングしたときのクラスタ数によるクラスタリング状態の評価値を演算装置で算出するクラスタリングの評価値算出方法において、クラスタリングされたデータに関する情報を取得する取得ステップと、取得した前記情報に基づいて前記評価値を算出する算出ステップとを有し、前記算出ステップは、クラスタ内のデータの分散の程度を示す各々の前記クラスタについての第１の指標値を当該クラスタのデータ数に基づく第１の値で規格化した値の各前記クラスタの総和である内的結合度と、前記クラスタ間の距離の指標となる各々の前記クラスタについての第２の指標値の総和を前記クラスタ数に基づく第２の値で規格化した外的分離度とをそれぞれ求め、前記内的結合度と前記外的分離度とを変数とする所定の演算式から前記評価値を算出するものである。

[0010] また、本発明のクラスタ数決定方法は、上記評価値算出方法を用いて、クラスタ数の異なるクラスタリングの結果に対する評価値をそれぞれ算出し、各前記評価値のうちの極値、または最大値若しくは最小値の評価値に対応するクラスタ数を最適値として判定するものである。

発明の効果

[0011] 本発明によれば、内的結合度をクラスタ内の各データの分散の程度を示す各々のクラスタについての第１の指標値をクラスタのデータ数に基づく第１の値で規格化した値の総和とし、外的分離度をクラスタ間の距離の指標となる各々のクラスタについての第２の指標値の総和をクラスタ数に基づく第２の値で規格化したものとして、評価値を算出するので、過剰な分類を抑制することができる評価値を得ることができる。

図面の簡単な説明

[0012] [図1]本発明を実施した演算装置の構成を示すブロック図である。

[図2]評価値が極大値となるクラスタリング結果を取得する処理手順を示すフローチャートである。

[図3]評価値が最大値となるクラスタリング結果を取得する処理手順を示すフローチャートである。

[図4]クラスタ間距離を用いた評価値Ｅ１（ｃ）を求める場合の各値を説明する説明図である。

[図5]基準点クラスタ間距離を用いた評価値Ｅ２（ｃ）を求める場合の各値を説明する説明図である。

[図6]サンプル１とデータの分類の状態を示す説明図である。

[図7]サンプル２を示す説明図である。

[図8]サンプル３を示す説明図である。

[図9]サンプル１のクラスタ数に対する評価値Ｅ１（ｃ）の変化を示すグラフである。

[図10]サンプル１のクラスタ数に対する評価値Ｅ２（ｃ）の変化を示すグラフである。

[図11]サンプル１のクラスタ数に対する評価値ＶＲＣ（ｃ）の変化を示すグラフである。

[図12]サンプル２のクラスタ数に対する評価値Ｅ１（ｃ）の変化を示すグラフである。

[図13]サンプル２のクラスタ数に対する評価値Ｅ２（ｃ）の変化を示すグラフである。

[図14]サンプル２のクラスタ数に対する評価値ＶＲＣ（ｃ）の変化を示すグラフである。

[図15]サンプル３のクラスタ数に対する評価値Ｅ１（ｃ）の変化を示すグラフである。

[図16]サンプル３のクラスタ数に対する評価値Ｅ２（ｃ）の変化を示すグラフである。

[図17]サンプル３のクラスタ数に対する評価値ＶＲＣ（ｃ）の変化を示すグラフである。

[図18]サンプルＡ～Ｅを示す説明図である。

発明を実施するための形態

[0013] 図１において、演算装置１０は、複数のデータに対してクラスタリングを行うとともに、クラスタリング状態を評価する評価値を算出し、その評価値に基づいて最適なクラスタリング数でのクラスタリング結果を取得するように構成される。最適なクラスタリング数を得るために、演算装置１０は、複数のデータに対して、クラスタ数を変えて複数回のクラスタリングを行う。演算装置１０は、例えばコンピュータで構成されており、評価値演算部１１、クラスタリング部１２、これらを統括的に制御する制御部１３、データ記憶部１４として機能する。

[0014] データ記憶部１４は、コンピュータのハードディスクやメモリで構成されており、クラスタリングの対象となる複数のデータが予め書き込まれている。これらデータは、ｑ（ｑは、１以上の整数）次元ベクトルであり、例えば画像の色や濃淡、色の分布等の内容をｑ次元ベクトルにしたものである。データ記憶部１４には、クラスタリング部１２で実施されたクラスタリング結果や、クラスタリングの演算途中の結果等が書き込まれる。

[0015] クラスタリング部１２は、データ記憶部１４からクラスタリングの対象となるデータを読み出して、クラスタリングを行い、その結果をデータ記憶部１４に保存する。クラスタリング部１２によるクラスタリングは、この例では上述のｋ－ｍｅａｎｓ（ｋ平均）法を用いて行われる。評価値演算部１１による評価値の算出手法は、ｋ－ｍｅａｎｓ法に限らず、各データを１つのクラスタに属するように分類するハードクラスタリングに適用することができる。このようなハードクラスタリングとしては、「ｋ－ｍｅａｎｓ＋＋」法、スペクトラルクラスタリング、単連結法、ウォード法等がある。

[0016] 評価値演算部１１は、クラスタリング部１２が１つのクラスタ数ｃ（ｃは、２以上の整数）についてクラスタリングを行うごとに、評価値Ｅ１（ｃ）を算出する。評価値演算部１１は、データ記憶部１４からクラスタリングされたデータに関する情報としてのクラスタリング結果を取得する。評価値演算部１１は、取得したクラスタリング結果に基づいて、そのクラスタリング

結果についてのクラスタリング状態を評価する評価値Ｅ１（ｃ）を算出する。評価値演算部１１は、評価値Ｅ１（ｃ）を制御部１３に出力する。なお、この評価値Ｅ１（ｃ）の算出の詳細については後述する。

[0017] 評価値演算部１１が取得するクラスタリングされたデータに関する情報としては、クラスタリングの演算途中に生成される結果、例えばｋ－ｍｅａｎｓ法では後述するようにクラスタの重心、クラスタ内距離等をクラスタリングの際に求めているので、それらを含めることができ、それを評価値Ｅ１（ｃ）の算出に利用することができる。このようにクラスタリングの際にされる演算結果を評価値Ｅ１（ｃ）の算出に利用することは演算時間の短縮等の点で好ましい。

[0018] 制御部１３は、評価値演算部１１で算出された各評価値Ｅ１（ｃ）を比べることにより、最適なクラスタ数ｃを判別する。この例では、制御部１３は、図２に示す山登り方式の手順に基づいて、極値（この例では極大値）をとる評価値Ｅ１（ｃ）に対応するクラスタ数ｃを最適なものと判別する。すなわち、評価値Ｅ１（ｃ）が算出されるごとに、その算出された評価値Ｅ１（ｃ）と前回に算出された評価値Ｅ１（ｃ）とを比べ、後者が前者よりも小さくなった場合に、前回に算出された評価値Ｅ１（ｃ）が極大値であるとして、その評価値Ｅ１（ｃ）に対応するクラスタ数ｃを最適なものと判別する。

[0019] 図２に示すように、制御部１３は、ステップＳＴ１でクラスタ数ｃを初期化するとともに、前回の評価値を示す変数Ｅ０を「０」にして初期化する。クラスタ数ｃは、この例の初期化では「２」に設定される。ステップＳＴ２で、制御部１３は、設定されたクラスタ数の下でｋ－ｍｅａｎｓ法によるクラスタリングをクラスタリング部１２に実施させ、その後のステップＳＴ３でクラスタリングの結果に基づいたクラスタ数ｃについての評価値Ｅ１（ｃ）を評価値演算部１１に算出させる。なお、クラスタ数ｃは、上述のｋ－ｍｅａｎｓ法におけるクラスタ数ｋである。

[0020] ステップＳＴ４では、制御部１３は、今回算出した評価値Ｅ１（ｃ）と変数Ｅ０とを比較する。この比較において、評価値Ｅ１（ｃ）が変数Ｅ０以上

である場合には、ステップＳＴ５、ＳＴ６の後に、再びステップＳＴ２においてクラスタリング部１２にクラスタリングを実施させる。ステップＳＴ５では、変数Ｅ０を今回算出された評価値Ｅ１（ｃ）にし、ステップＳＴ６では、クラスタ数ｃを「１」だけインクリメントする。これにより、クラスタ数ｃを「１」だけ増加させて次のクラスタリングが行われ（ステップＳＴ２）、ステップＳＴ３で新たに得られる評価値Ｅ１（ｃ）が前回のものと比較される（ステップＳＴ４）。

[0021] 一方、評価値Ｅ１（ｃ）が変数Ｅ０よりも小さい場合には、制御部１３は、評価値Ｅ１（ｃ）が減少に転じたものとして、ステップＳＴ７により、クラスタ数（ｃ－１）に対応したクラスタリング結果を取得する。すなわち、クラスタリングを最後に行ったときのクラスタ数をｃとしたときのクラスタ数（ｃ－１）がクラスタ数として最適な値であると判定する。

[0022] なお、設定されたクラスタ数ｃの範囲内で最大となるクラスタ数を最適な値として判別してもよい。この場合は、図３に手順を示すように、まずステップＳＴ１１でクラスタ数ｃの初期値と設定最大値が設定されることで、評価するクラスタ数ｃの範囲が設定される。制御部１３は、ステップＳＴ１２でクラスタ数ｃを初期値に設定する。ステップＳＴ１３では、ステップＳＴ１２で設定されたクラスタ数ｃの下でクラスタリングをクラスタリング部１２で行い、その後のステップＳＴ１４でクラスタリングの結果に基づいたクラスタ数ｃについての評価値Ｅ１（ｃ）を評価値演算部１１に算出させる。ステップＳＴ１５では、クラスタ数ｃが設定された設定最大値以上であるか否かを調べる。クラスタ数ｃが設定最大値に達していない場合には、ステップＳＴ１６でクラスタ数ｃを「１」だけインクリメントした後に、再びステップＳＴ１３によるクラスタリングを行う。これにより、クラスタ数ｃが設定最大値に達するまで、クラスタ数ｃを「１」ずつインクリメントしながら繰り返しクラスタリングが行われ、クラスタリングごとに評価値Ｅ１（ｃ）が算出される。そして、設定最大値のクラスタ数ｃについてのクラスタリングの終了後、各クラスタ数ｃについての評価値Ｅ１（ｃ）が比較され、その

中で最大の評価値Ｅ１（ｃ）に対応するクラスタ数ｃを最適な値と判定し、そのクラスタ数ｃに対応したクラスタリング結果を制御部１３が取得する（ステップＳＴ１７）。

[0023] 次に、評価値演算部１１が算出する評価値について説明する。評価値演算部１１が算出する評価値Ｅ１（ｃ）は、式（１）のように表される。

[数1]

$$El(c)=\frac{\frac{\sum_{i=1}^{c}\min_{i\neq j} D(v_i,v_j)}{c}}{\sum_{i=1}^{c}\frac{1}{n_i}\sum_{x\in c_i} D(x,v_i)} \quad \cdots (1)$$

但し、 $D(v_i,v_j)=dist(v_i,v_j)^2$

$D(x,v_i)=dist(x,v_i)^2$

[0024] 式（１）中の各値は、次の通りである。なお、ｉ＝１、２、・・・ｃである。また、クラスタ重心は、クラスタ内の各データの相加平均として求められる。図４には、データｘ、ｉ番目のクラスタc_i、クラスタc_iの重心v_i、クラスタ間距離ｄｉｓｔ（v_i，v_j）、クラスタ内距離ｄｉｓｔ（ｘ、v_i）の一例を模式的に示す。

ｎ：クラスタリングの対象となるデータ数（全データ数）

ｘ：データ

ｃ：クラスタ数

c_i：ｉ番目のクラスタ

n_i：クラスタc_iのデータ数

v_i：クラスタc_iの重心（クラスタ重心）

Ｄ（ｘ、v_i）：クラスタc_i内のデータｘについてのクラスタ内距離指標値

Ｄ（v_i，v_j）：クラスタc_iと他のクラスタc_jとのクラスタ間距離指標値

ｄｉｓｔ（ｘ、ｖ$_i$）：クラスタｃ$_i$内のクラスタ重心とデータｘとの距離（クラスタ内距離）

ｄｉｓｔ（ｖ$_i$，ｖ$_j$）：クラスタｃ$_i$と他のクラスタｃ$_j$とのクラスタ重心間の距離（クラスタ間距離）

[0025] 式（１）の右辺の分母は、クラスタ内におけるデータの分散（データ同士の類似）の程度を全クラスタについて示す内的結合度になっている。クラスタ内距離指標値（＝Ｄ（ｘ、ｖ$_i$））は、クラスタ内に設定される距離の基点としての第１の代表点とクラスタ内のデータとの距離の程度（隔たりの大きさ）を示すものであり、クラスタ内の各々のデータについて求められる。この例では、第１の代表点をクラスタ重心とし、このクラスタ重心とデータとの距離であるクラスタ内距離（＝ｄｉｓｔ（ｘ、ｖ$_i$））の二乗の値をクラスタ内距離指標値にしている。

[0026] 内的結合度は、クラスタごとの第１の指標値を当該クラスタ内のデータ数で除することで規格化した値の各クラスタについての総和として求められる。各クラスタの第１の指標値は、対応するクラスタ内のデータの分散の程度を示しており、この例では、対応するクラスタ内の各データについてのクラスタ内距離指標値の和である。

[0027] 式（１）の右辺の分子は、クラスタの分離の程度を全クラスタについて示す外的分離度になっている。クラスタ間距離指標値（＝Ｄ（ｖ$_i$，ｖ$_j$））は、１のクラスタと他の１のクラスタとの間の距離の程度を示しており、この例ではクラスタ間の距離の基点となる各クラスタの第２の代表点をクラスタ重心として、１のクラスタと他の１のクラスタとのクラスタ重心間の距離であるクラスタ間距離（＝ｄｉｓｔ（ｖｉ，ｖｊ））の二乗の値である。

[0028] 外的分離度は、クラスタごとの第２の指標値の各クラスタについての総和をクラスタ数で除して規格化した値である。第２の指標値は、１のクラスタと他の各クラスタとの分離の程度、すなわち当該１のクラスタについてのクラスタ間の距離の指標となる値であり、この例では、各々のクラスタについてのクラスタ間距離指標値の最小値である。

[0029] また、この例では、クラスタ内距離、クラスタ間距離をいずれもユークリッド距離としている。２つのｑ次元ベクトルであるデータをそれぞれａ（a_1，a_2，・・・a_q）、ｂ（b_1，b_2，・・・b_q）とすると、ａとｂとのユークリッド距離は、式（Ａ）のように表される。

[0030] [数2]

$$\mathrm{dist}(a,b)=\|a-b\|=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_q-b_q)^2} \quad \cdots (A)$$

[0031] 評価値Ｅ１（ｃ）では、上記のように、外的分離度は、クラスタごとのクラスタの分離の程度の指標となる第２の指標値の各クラスタについての総和をクラスタ数で規格化した値であり、内的結合度は、クラスタごとのクラスタ内の各データの分散の程度を示す第１の指標値をクラスタ内のデータ数で規格化した値の各クラスタについての総和である。このため、全データ内において、他のデータのまとまりに対して、範囲の大きなデータのまとまりがある場合やデータの分布密度が高いデータのまとまりがある場合等であっても、個々のクラスタ内におけるデータの分散の程度が内的結合度に適正に反映される。すなわち、過剰な分類がされた場合に、評価値Ｅ１（ｃ）がより大きくなるようなことがなく過剰な分類が抑制される。

[0032] 評価値Ｅ１（ｃ）は、クラスタリング状態が良好であれば、その程度に応じて値が大きくなる。したがって、評価値Ｅ１（ｃ）が、極大値または最大値をとるようにクラスタ数ｃを決めればよい。なお、式（１）の右辺の逆数を評価値Ｅ１（ｃ）とする場合には、評価値Ｅ１（ｃ）が、極小値または最小値をとるようにクラスタ数ｃを決めればよい。また、クラスタ重心やクラスタ間距離は、ｋ－ｍｅａｎｓ法の演算過程の途中結果を用いることができ演算数の削減に有利である。

[0033] 上記評価値Ｅ１（ｃ）では、クラスタ間の距離の指標である第２の指標値をクラスタ間距離指標値を用いて求めているが、式（２）に示す評価値Ｅ２

（ｃ）のように、予め設定された基準点からのクラスタまでの距離の程度を示す基準点クラスタ間距離指標値を用いて求めた値を第２の指標値とすることができる。評価値Ｅ２（ｃ）は、詳細を後述するように、それを求める際の演算数が評価値Ｅ１（ｃ）に比べて少なくなる。

[0034] [数3]

$$E2(c)=\frac{\frac{\sum_{i=1}^{c} n_i D(v_i,m)}{c}}{\sum_{i=1}^{c}\frac{1}{n_i}\sum_{x\in c_i} D(x,v_i)} \quad \cdots (2)$$

但し、 $D(v_i,m)=dist(v_i,m)^2$

$D(x,v_i)=dist(x,v_i)^2$

[0035] 式（２）中の各値は、次の通りである。なお、その他の値は、式（１）のものと同じである、また、この例では、基準点をクラスタリングの対象となるｎ個の全データの重心としている。図５には、データｘ、ｉ番目のクラスタc_i、クラスタc_iの重心v_i、基準点クラスタ間距離ｄｉｓｔ（v_i，ｍ）、クラスタ内距離ｄｉｓｔ（ｘ、v_i）の一例を模式的に示す。

ｍ：基準点

Ｄ（ｖｉ，ｍ）：クラスタc_iの基準点クラスタ間距離指標値

ｄｉｓｔ（ｖｉ，ｍ）：クラスタc_iのクラスタ重心と基準点との距離（基準点クラスタ間距離）

[0036] 上記式（２）は、評価値Ｅ１（ｃ）の式（１）と同じく、その右辺の分子が外的分離度であり、右辺の分母が内的結合度である。評価値Ｅ２（ｃ）の外的分離度は、第２の指標値の各クラスタにつての総和であるが、基準点クラスタ間距離指標値をクラスタ内のデータ数で重み付けした値を第２の指標値としている。基準点クラスタ間距離指標値は、この例では、クラスタリング対象となるｎ個の全データの重心（以下、データ重心という）を基準点とし、基準点に対するクラスタリング側の距離の基点となる第２の代表点をク

ラスタ重心として、基準点とクラスタ重心との距離である基準点クラスタ間距離（＝ｄｉｓｔ（ｖｉ，ｍ））の二乗の値を基準点クラスタ間距離指標値としている。なお、基準点クラスタ間距離指標値に重み付けをする場合、重み付けする値は、クラスタ内のデータ数に限定されず、クラスタ内のデータ数に基づく値で重み付けしてもよい。

[0037] 上記の評価値Ｅ２（ｃ）においても、評価値Ｅ１（ｃ）と同様に、外的分離度が、クラスタ間の距離の指標となるクラスタごとの第２の指標値の各クラスタについての総和をクラスタ数で規格化した値であり、内的結合度が、クラスタごとのクラスタ内の各データの分散の程度を示す第１の指標値をクラスタ内のデータ数で規格化した値の各クラスタについての総和になっているから、過剰な分類がされた場合に、評価値Ｅ２（ｃ）がより大きくなるようなことがなく過剰な分類が抑制される。

[0038] 評価値Ｅ２（ｃ）は、評価値Ｅ１（ｃ）と同じく、クラスタリング状態が良好であれば、その程度に応じて値が大きくなる。したがって、評価値Ｅ２（ｃ）が、極大値または最大値をとるようにクラスタ数ｃを決めればよい。なお、式（２）の右辺の逆数を評価値Ｅ２（ｃ）とする場合には、評価値Ｅ２（ｃ）が、極小値または最小値をとるようにクラスタ数ｃを決めればよい。

[0039] なお、上記の評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）のように、クラスタ内距離指標値、クラスタ間距離指標値、基準点クラスタ間距離指標値として、対応するユークリッド距離（クラスタ内距離、クラスタ間距離、基準点クラスタ間距離）を二乗した値を用いる他に、対応するユークリッド距離そのものを用いることもできる。上記の例のように、対応するユークリッド距離を二乗した値を用いる場合には、ユークリッド距離そのものを用いるよりも、計算量を少なくすることができるという利点がある。

[0040] 上記評価値Ｅ１（ｃ）及び評価値Ｅ２（ｃ）におけるクラスタ内距離、クラスタ間距離、基準点クラスタ間距離は、ユークリッド距離に限定されない。例えば、式（Ｂ）に示すマンハッタン距離や式（Ｃ）に示すミンコフスキ

一距離であってもよい。なお、式（Ｃ）におけるｐは１以上の任意の値であるが、周知のようにｐ＝１のときはマンハッタン距離に、ｐ＝２のときはユークリッド距離と同義である。また、この他にも「Point symmetry distance」等を用いてもよい。

[0041] [数4]

$$dist(a,b)=|a_1-b_1|+|a_2-b_2|+\cdots+|a_q-b_q| \quad \cdots (B)$$

[0042] [数5]

$$dist(a,b)=\left(\sum_{i=1}^{q}|a_i-b_i|^p\right)^{\frac{1}{p}} \quad \cdots (C)$$

[0043] さらに、評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）では、クラスタ内の各データについてのクラスタ内距離の基点である第１の代表点、クラスタ間距離の各クラスタの基点または基準点クラスタ間距離のクラスタ側の基点である第２の代表点をいずれもクラスタ重心にしているが、これに限定されない。例えば、第１の代表点、第２の代表点を各クラスタ内において、クラスタ重心に最も近いデータとしてもよい。また、第２の代表点は、クラスタ同士または基準点とクラスタとの離れた具合（距離）を測る上での基点であるから、隣接したクラスタあるいは基準点に最も近いまたは最も離れたクラスタ内のデータ等としてもよい。さらに、第１の代表点及び第２の代表点を、別途定めた基準によって決まるクラスタ内の点またはデータやクラスタ内の任意の点またはデータとしてもよい。

[0044] 式（２）の基準点は、任意の点やデータに設定することができ、上記のようなデータ重心の他に、データ重心に最も近いデータ、いずれかのクラスタの第２の代表点、外部より指定する点やデータ、外部から指定された記憶装置のアドレスに格納されているデータ、最初に入力されたデータ等のように別途定めた基準によって決まる点やデータにしてもよい。

[0045] また、評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）では、内的結合度を求める際

に、第１の指標値を、第１の値としての当該クラスタのデータ数n_iで規格化しているが、規格化するための第１の値は、これに限定されるものではなく、当該クラスタのデータ数n_iに基づく値とすることができる。例えば、当該クラスタのデータ数n_iの冪（n_i^2、n_i^3、$n_i^{1/2}$、n_i^{-2}等）、データ数n_iを定数倍した値、データ数n_iに一定数を減算または加算した値、データ数n_iとクラスタ数ｃとを予め決めた関数Ｆ（n_i，ｃ）、例えば「Ｆ（n_i，ｃ）＝n_i－ｃ」、ｋを定数とした「Ｆ（n_i，ｃ）＝ｋ・n_i－ｃ」等に適用した結果等を第１の値として用いることができる。

[0046] また、外的分離度を求める際に、第２の値としてのクラスタ数ｃで規格化しているが、第２の値は、これに限定されるものではなく、クラスタ数ｃに基づく値で規格化することができる。クラスタ数ｃ以外のクラスタ数に基づく値としては、第１の値と同様に、例えば、クラスタ数ｃの冪、クラスタ数ｃを定数倍した値、クラスタ数ｃに一定数を減算または加算した値、クラスタ数ｃと全データ数ｎとを予め決めた関数Ｇ（ｎ，ｃ）に適用した結果等を第２の値として用いることができる。

[0047] 上記では、評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）は、評価値を求める演算を内的結合度と外的分離度との比とするものであるが、評価値を求める演算はこれに限定されない。上記のように求められる内的結合度は、その値が小さいほど、クラスタ内のデータ相互の類似性が高くなるように各クラスタ内のデータがまとまったクラスタリング状態となることを意味する。一方で、上記のように求められる外的分離度は、その値が大きいほど、クラスタ相互間の類似性が低くなるように各クラスタが相互に離れたクラスタリング状態となることを意味する。このため、内的結合度と外的分離度とから評価値を求める演算は、内的結合度が小さくなる方向に変化したとき、また外的分離度が大きくなる方向に変化したときに、評価値が増加または減少の一方に変化するようにすればよく、そのような結果が得られる内的結合度と外的分離度とを変数とする演算式を用いればよい。

[0048] 評価値を求める演算としては、上述のように内的結合度と外的分離度との

比とするものの他に、例えば、次の各式に示すように、内的結合度と外的分離度の一方の逆数と他方とをそれぞれ重み付けして加算することで評価値Ｅを求めるもの、内的結合度と外的分離度の一方から他方をそれぞれ重み付けして減算することで評価値Ｅを求めるもの等を用いることができる。なお、式中の値Compは内的結合度を、Sepは外的分離度を、Ｗａ、Ｗｂは重み付けの定数であり、Ｗａ，Ｗｂ≠０である。

Ｅ＝Ｗｂ・Ｓｅｐ＋Ｗａ・（１／Ｃｏｍｐ）

Ｅ＝Ｗｂ・（１／Ｓｅｐ）＋Ｗａ・Ｃｏｍｐ

Ｅ＝Ｗｂ・Ｓｅｐ－Ｗａ・Ｃｏｍｐ

Ｅ＝Ｗａ・Ｃｏｍｐ－Ｗｂ・Ｓｅｐ

実施例

[0049] 以下、サンプルのデータをクラスタリングし、評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）を用いて最適なクラスタ数を決定した実施例１、２について説明する。実施例１、２では、クラスタリング手法としてｋ－ｍｅａｎｓ法を用い、最適なクラスタ数の決定の手順としては、図３に示すものを用いて評価値が最大となるものとした。クラスタリング数の範囲は、２～１０とした。また、比較例１として、同様な条件の下で、非特許文献１に記載された評価値ＶＲＣ（ｃ）を用いて最適なクラスタ数を決定した。

[0050] 評価値ＶＲＣ（ｃ）の算出式を式（３）に示す。なお、式（３）の各値は、式（１）または式（２）のもとの同じである。評価値ＶＲＣ（ｃ）では、式（６）の分母の値が内的結合度を示しており、分子の値が外的分離度を示している。

[0051]

[数6]

$$VRC(c) = \frac{\dfrac{\sum_{i=1}^{c} n_i D(v_i,m)}{c-1}}{\dfrac{\sum_{i=1}^{c} \sum_{x \in c_i} D(x,vi)}{n-c}} \quad \cdots (3)$$

但し、 $D(v_i,m) = dist(v_i,m)^2$

$D(x,v_i) = dist(x,v_i)^2$

[0052] 実施例１、２及び比較例１では、図６（ａ）に示すサンプル１、図７に示すサンプル２、図８に示すサンプル３に対してクラスタリングを実施した。これら図６（ａ）、図７、図８は、いずれも２次元ベクトルのデータの分布状態を示している。

[0053] サンプル１では、データが互いに離れた小領域と、この小領域よりもサイズが大きい大領域にそれぞれ同じ密度で分布させている。小領域のデータ数は、１００個、大領域のデータ数は、４００個であり、大領域は、小領域の４倍の面積である。このサンプル１について、クラスタ数ｃを２～１０の範囲で変化させてクラスタリングした場合の評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）及び評価値ＶＲＣ（ｃ）の変化を図９ないし図１１にそれぞれ示す。

[0054] 評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）のいずれの場合にも、クラスタ数ｃが「２」のときに評価値が最大となり、そのときのクラスタリングの状態は、図６（ｂ）に示すようなものになり、小領域のデータと大領域のデータとがそれぞれ１つのクラスタに分類される結果となった。この結果は、サンプル１のデータの分布に対する人間の感覚的な判断と一致している。一方、評価値ＶＲＣ（ｃ）では、クラスタ数ｃが「５」のときに評価値が最大となり、そのときのクラスタリングの状態は、図６（ｃ）に示すように、小領域のデータが１つのクラスタに分類され、大領域のデータは４つのクラスタに分類される結果になった。なお、クラスタリング手法として同じ手法を用いて

いるため、クラスタ数ｃが同じであれば、評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）、評価値ＶＲＣ（ｃ）のいずれにおいてもクラスタリングの状態は同じである。

[0055] 上記同様にして、サンプル２についての評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）、評価値ＶＲＣ（ｃ）の変化を調べた結果を図１２ないし図１４に、またサンプル３についての評価値Ｅ１（ｃ）、評価値Ｅ２（ｃ）、評価値ＶＲＣ（ｃ）の変化を調べた結果を図１５ないし図１７にそれぞれ示す。

[0056] サンプル２は、互いに離れた同じサイズの低密度領域と高密度領域にデータを分布させたものであり、低密度領域よりも高密度領域のデータの密度が高くなっている。サンプル２では、低密度領域のデータ数が５０個であり、高密度領域のデータ数が５００個である。サンプル３では、互いに離れた小領域とこの小領域よりもサイズが大きい大領域とに、異なる密度でデータを分布させたものであり、小領域よりも大領域のデータの密度が高くなっている。サンプル３では、データ数は、小領域が５０個、大領域が４００個であり、大領域は、小領域の４倍の面積である。

[0057] 評価値Ｅ１（ｃ）、Ｅ２（ｃ）は、上記サンプル２、サンプル３についても、サンプル１と同じく、クラスタ数ｃが「２」のときに最大となり、サンプル２、３の領域ごとにそれらのデータが１つのクラスタに分類される結果となった。一方、評価値ＶＲＣ（ｃ）は、サンプル２、サンプル３についても、クラスタ数ｃが「５」のときに最大となり、サンプル２の低密度領域、サンプル３の小領域のデータがそれぞれ１つのクラスタに分類され、サンプル２の高密度領域、サンプル３の大領域のデータがそれぞれ４つのクラスタに分類された。

[0058] 表１に、サンプル１～３についてのクラスタ数ｃが「２」と「５」の場合の各評価値の値を示す。なお、最大値となった評価値に対して表１中の「結果」の欄に丸印を付してある。

[0059]

[表1]

		クラスタ数c	サンプル1		サンプル2		サンプル3	
			評価値	結果	評価値	結果	評価値	結果
実施例1	E1(c)	2	2.944	○	6.165	○	4.177	○
		5	1.542		2.325		1.727	
実施例2	E2(c)	2	287.6	○	280.2	○	318.8	○
		5	176.5		153.3		179.7	
比較例1	VRC(c)	2	829.0		1102.4		860.3	
		5	1088.1	○	1194.9	○	1108.2	○

[0060] また、表２には、サンプル１について、クラスタ数ｃが「２」、「５」各場合の評価値Ｅ２（ｃ）、ＶＲＣ（ｃ）における内的結合度及び外的分離度の詳細を示す。クラスタ数ｃが「５」のときの評価値Ｅ２（ｃ）における内的結合度は、クラスタ数ｃが「２」のときのものに比べて僅かに小さな値になっているが、クラスタ数ｃが「５」のときの評価値ＶＲＣ（ｃ）における内的結合度は、クラスタ数ｃが「２」のものに比べて１／３以下の値であり

、かなり小さな値になっている。

[0061] [表2]

	実施例2(E2(c))		比較例1(VRC(c))	
クラスタ数c	2	5	2	5
内的結合度	0.8293	0.7766	0.5753	0.1575
外的分離度	238.4712	137.1031	476.9424	171.3788
評価値	287.5572	176.5428	829.0325	1088.1194

[0062] 上述のように評価値Ｅ２（ｃ）は、クラスタごとのクラスタ内距離指標値の和、すなわちクラスタごとの第１の指標値を、当該クラスタ内のデータ数n_iで除して規格化した値の各クラスタについての総和を内的結合度にしている。一方、評価値ＶＲＣ（ｃ）の内的結合度は、上記式（３）から分かるように、クラスタごとに求めた第１の指標値の各クラスタについての総和を「ｎ－ｃ（＝全データ数－クラスタ数）」で除して規格化した値である。なお、全データ数ｎに対してクラスタ数ｃの変化量ないし値を無視しても問題ないので、便宜上、全データ数ｎで除して規格化しているものとして説明する。

[0063] 上述のように、サンプル１は、クラスタ数ｃが「２」のときには、小領域のデータと大領域のデータとがそれぞれ１つのクラスタに分類され（図６（ｂ））、クラスタ数ｃが「５」のときには、小領域のデータが１つのクラスタに分類され、大領域のデータは４つのクラスタに分類される（図６（ｃ））。クラスタ数ｃが「２」のときに比べて、クラスタ数ｃが「５」ときには、大領域のデータは４つのクラスタに分類されることにともない、それら４個のクラスタ内の各データについてのクラスタ内距離指標値がかなり小さくなる。この結果、それら４個のクラスタのそれぞれについての第１の指標値（＝クラスタ内距離指標値の和）もかなり小さくなり、第１の指標値の各ク

ラスタについての総和もかなり小さくなる。

[0064] 評価値ＶＲＣ（ｃ）の内的結合度は、第１の指標値の各クラスタについての総和を全データ数ｎで規格化した値であり、全データ数ｎは変化しない値である。したがって、クラスタ数ｃが大きくなって第１の指標値の総和がかなり小さくなると、それに応じて内的結合度もかなり小さくなる。このため、クラスタ数ｃが「５」のときの評価値ＶＲＣ（ｃ）における内的結合度は、クラスタ数ｃが「２」のものに比べて１／３以下の値になっている。

[0065] 上記のようにクラスタ数ｃが大きくなることによって、第１の指標値がかなり小さくなる点については、評価値Ｅ２（ｃ）を求める場合においても全く同じである。しかしながら、評価値Ｅ２（ｃ）では、上記のように、クラスタごとの第１の指標値を当該クラスタ内のデータ数n_iで除して規格化しており、そのデータ数n_iは、クラスタの範囲が狭くなることによって小さな値になる。したがって、クラスタ数ｃが大きくなることによって第１の指標値が小さくなっても、内的結合度の変化は小さい。このため、上記のようにクラスタ数ｃが「５」のときの評価値Ｅ２（ｃ）における内的結合度は、クラスタ数ｃが「２」のものに比べて僅かな減少になっている。

[0066] クラスタ数の増減にともなう外的分離度の変化によっても、評価値Ｅ２（ｃ）、評価値ＶＲＣ（ｃ）は変化するが、サンプル１のようなデータの分布については、評価値ＶＲＣ（ｃ）では、上記のようにクラスタ数ｃが増大したときに内的結合度が大きく減少する傾向により、クラスタ数ｃが「５」のときに、その値が最も高くなり、人間の感覚的な判断よりも過剰に分類された結果になったと考えられる。これに対して、評価値Ｅ２（ｃ）では、クラスタ数ｃが増大したときに内的結合度が大きく減少する傾向がないため、クラスタ数ｃが「２」のときに最も高くなり、過剰な分類が抑制されて人間の感覚的な判断と一致した結果になったと考えられる。なお、他のサンプル２、３についてもサンプル１と同様であり、評価値Ｅ１（ｃ）についても評価値Ｅ２（ｃ）と同様である。

[0067] また、図１８に示すサンプルＡ～Ｅについて、実施例３、比較例２、３を

実施した。実施例３では評価値Ｅ１（ｃ）を、比較例２では評価値ＶＲＣ（ｃ）を、比較例３では、既知のＤＢインデックス（以下、評価値ＤＢ（ｃ）と称する）をそれぞれ用いて、最適なクラスタ数を調べた。サンプルＡ～Ｅのうちサンプルｂは、データが三次元ベクトルであり、その他はデータが二次元ベクトルである。評価値ＤＢ（ｃ）は、式（４）で求められる。

[0068] [数7]

$$DB(c) = \frac{1}{c}\sum_{i=1}^{c}\max_{j\neq i}\{D_{i,j}\} \quad \cdots (4)$$

但し、$D_{i,j} = \frac{(\bar{d}_i+\bar{d}_j)}{d_{i,j}}$

[0069] 式（４）中の各値は、次の通りである。

$\bar{d}_i$（上にバー）：ｉ番目のクラスタにおけるクラスタ内距離の平均値

$\bar{d}_j$（上にバー）：ｊ番目のクラスタにおけるクラスタ内距離の平均値

$d_{i,j}$：ｉ番目とｊ番目の各クラスタの重心間の距離（クラスタ間距離）

[0070] 上述のように、評価値Ｅ１（ｃ）、Ｅ２（ｃ）では、クラスタ内の各データの分散の程度を示す各々のクラスタについての第１の指標値をクラスタのデータ数に基づく第１の値で規格化した値の総和を内的結合度とし、クラスタ間の距離の指標となる各々のクラスタについての第２の指標値の総和をクラスタ数に基づく第２の値で規格化した値を外的分離度としており、評価値Ｅ１（ｃ）、Ｅ２（ｃ）は、それら内的結合度と外的分離度との比として求められる。これに対して、評価値ＤＢ（ｃ）では、一のクラスタを含む一対のクラスタに関してクラスタリング状態が最も悪いと評価される他の一のクラスタとの組み合わせにおける各クラスタ内のデータ分散の程度と当該一のクラスタ及び当該の他の一のクラスタの分離の程度との比を各クラスタについてそれぞれ求め、評価値ＤＢ（ｃ）は、それら各クラスタの比の平均であり、評価値Ｅ１（ｃ）、Ｅ２（ｃ）とは異なる。

[0071] 各評価値を用いて判別された最適なクラスタ数の結果を表３に示す。なお

、評価値Ｅ１（ｃ）、評価値ＶＲＣ（ｃ）については、それらが最大値となるクラスタ数を最適なクラスタ数とし、評価値ＤＢ（ｃ）は、それが最小値となるクラスタ数を最適なクラスタ数とした。

[0072] [表3]

		サンプル				
		A	B	C	D	E
実施例3	E1(c)	10	4	5	5	3
比較例2	VRC(c)	10	4	11	8	6
比較例3	DB(c)	10	4	2	5	3

[0073] 実施例３では、サンプルＡ～Ｅのいずれにおいても、クラスタリング数を含むクラスタリングの結果が人間の感覚的な判断と一致した結果となった。これに対して、比較例２は、サンプルＣ～Ｅにおいて、人間の感覚的な判断によるクラスタ数と比較してクラスタ数が多くなり過剰に分割される結果となった。また、比較例３では、人間の感覚的な判断と比較して、比較例２のように過剰な分割とはならなかったが、サンプルＣにおいてクラスタリング数が少ない結果となった。すなわち、比較例３の評価値ＤＢ（ｃ）は、データの複数のまとまりが相互に近接している場合には、それらをまとめて１つのクラスタに分類したときに値が高くなる傾向があることが分かる。

符号の説明

[0074] １０　演算装置

１１　評価値演算部

１２　クラスタリング部

１３　制御部

１４　データ記憶部

Ｅ１（ｃ）　評価値

Ｅ２（ｃ）　評価値

請求の範囲

[請求項1] 複数のデータをクラスタリングしたときのクラスタ数によるクラスタリング状態の評価値を演算装置で算出するクラスタリングの評価値算出方法において、

クラスタリングされたデータに関する情報を取得する取得ステップと、

取得した前記情報に基づいて前記評価値を算出する算出ステップとを有し、

前記算出ステップは、

クラスタ内のデータの分散の程度を示す各々の前記クラスタについての第１の指標値を当該クラスタのデータ数に基づく第１の値で規格化した値の各前記クラスタの総和である内的結合度と、

前記クラスタ間の距離の指標となる各々の前記クラスタについての第２の指標値の総和を前記クラスタ数に基づく第２の値で規格化した外的分離度と

をそれぞれ求め、

前記内的結合度と前記外的分離度とを変数とする所定の演算式から前記評価値を算出する

ことを特徴とするクラスタリングの評価値算出方法。

[請求項2] 各々の前記クラスタの前記第１の指標値は、当該クラスタ内の第１の代表点と当該クラスタ内の各データとの間のクラスタ内距離の和または二乗和であることを特徴とする請求項１に記載のクラスタリングの評価値算出方法。

[請求項3] 前記第１の代表点は、前記クラスタ内のデータの重心であることを特徴とする請求項２に記載のクラスタリングの評価値算出方法。

[請求項4] 前記第１の値は、前記クラスタ内のデータ数であることを特徴とする請求項１ないし３のいずれか１項に記載のクラスタリングの評価値算出方法。

[請求項5] 各々の前記クラスタの前記第２の指標値は、当該クラスタ内の第２の代表点と他の前記クラスタ内の第２の代表点との間のクラスタ間距離のうち、最も小さい前記クラスタ間距離または最も小さい前記クラスタ間距離を二乗した値であることを特徴とする請求項１ないし４のいずれか１項に記載のクラスタリングの評価値算出方法。

[請求項6] 各々の前記クラスタの前記第２の指標値は、全データの基準点と当該クラスタ内の第２の代表点との間の基準点クラスタ間距離または前記基準点クラスタ間距離を二乗した値であることを特徴とする請求項１ないし４のいずれか１項に記載のクラスタリングの評価値算出方法。

[請求項7] 各々の前記クラスタの前記第２の指標値は、全データの基準点と当該クラスタ内の第２の代表点との間の基準点クラスタ間距離または前記基準点クラスタ間距離を二乗した値のいずれかに当該クラスタ内のデータ数に基づく重み付けをした値であることを特徴とする請求項１ないし４のいずれか１項に記載のクラスタリングの評価値算出方法。

[請求項8] 前記基準点は、全データの重心であることを特徴とする請求項６または７に記載のクラスタリングの評価値算出方法。

[請求項9] 前記第２の代表点は、前記クラスタ内のデータの重心であることを特徴とする請求項６ないし８のいずれか１項に記載のクラスタリングの評価値算出方法。

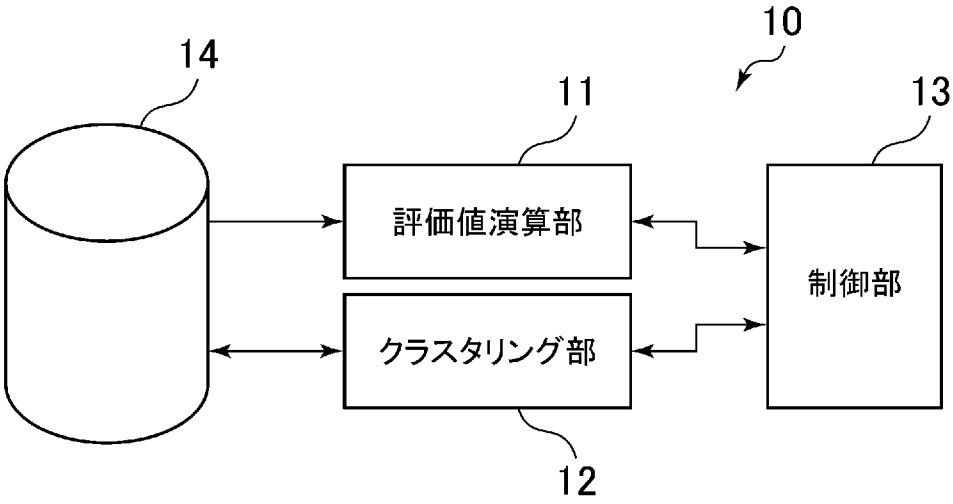
[請求項10] 前記内的結合度と前記外的分離度との比を前記評価値として算出することを特徴とする請求項１ないし９のいずれか１項に記載のクラスタリングの評価値算出方法。

[請求項11] 請求項１ないし１０のいずれか１項に記載のクラスタリングの評価値算出方法を用いて、前記クラスタ数の異なるクラスタリングの結果に対する前記評価値をそれぞれ算出し、

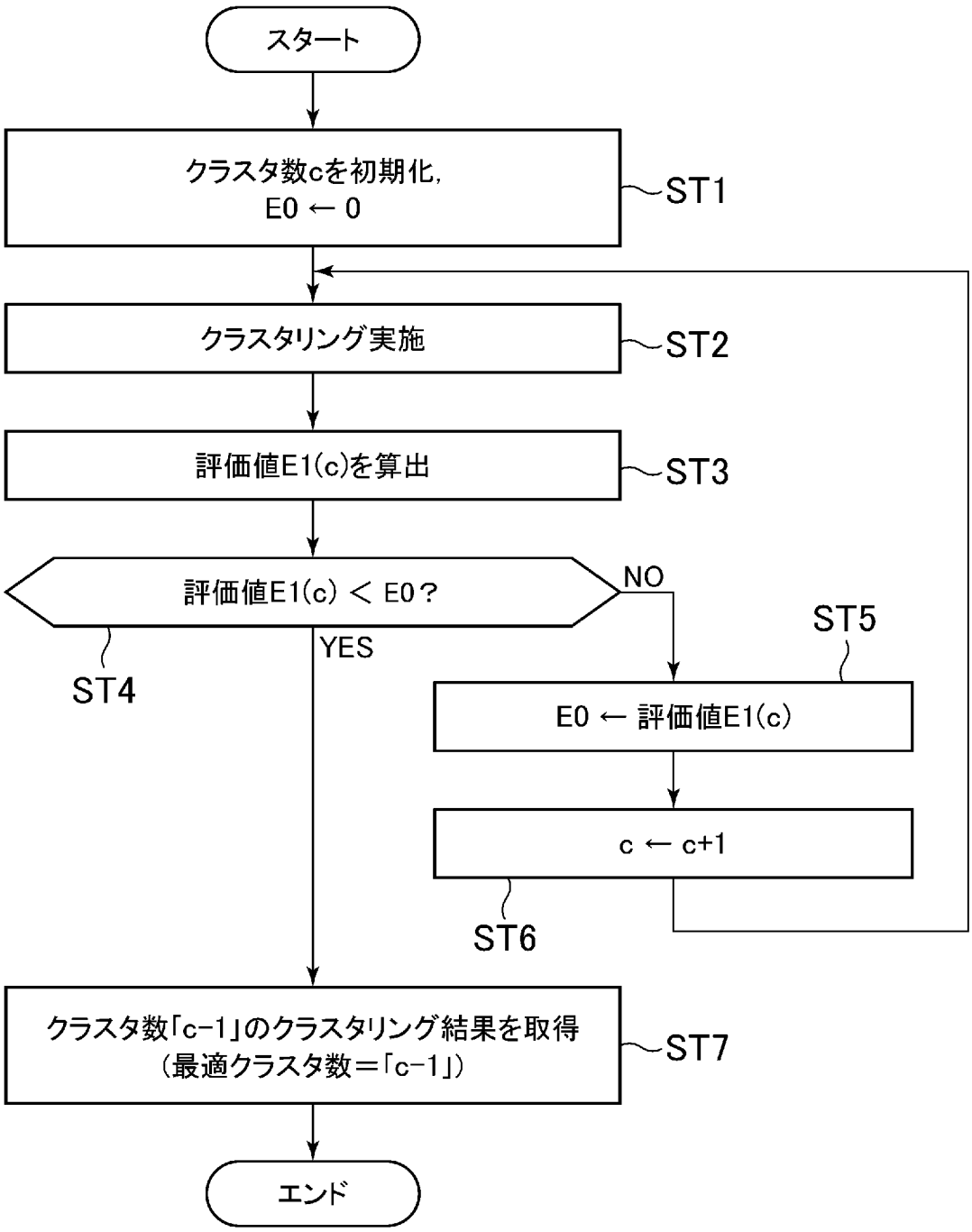
各前記評価値のうちの極値、または最大値若しくは最小値の前記評価値に対応する前記クラスタ数を最適値として判定する

ことを特徴とするクラスタリング数決定方法。

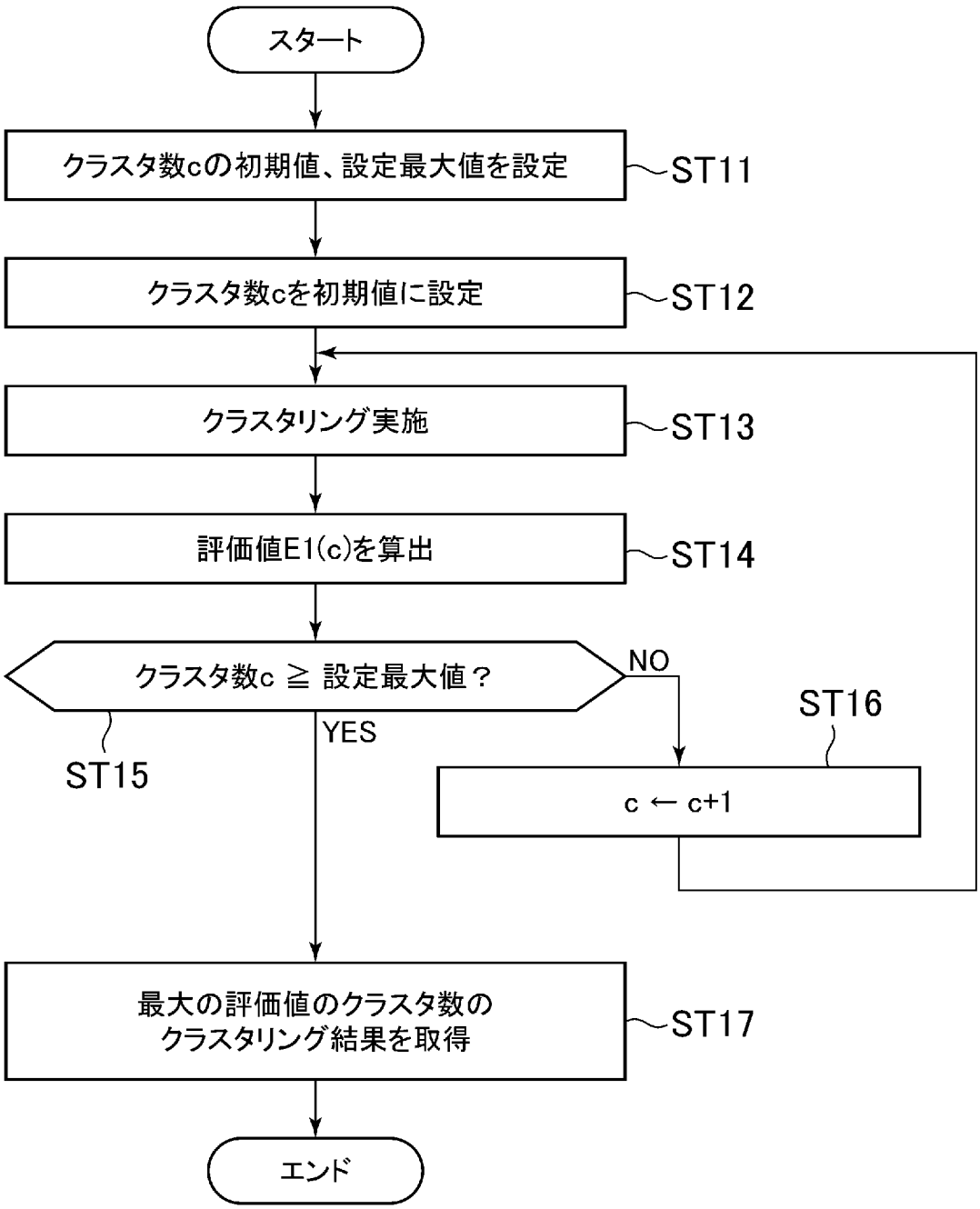
[図1]



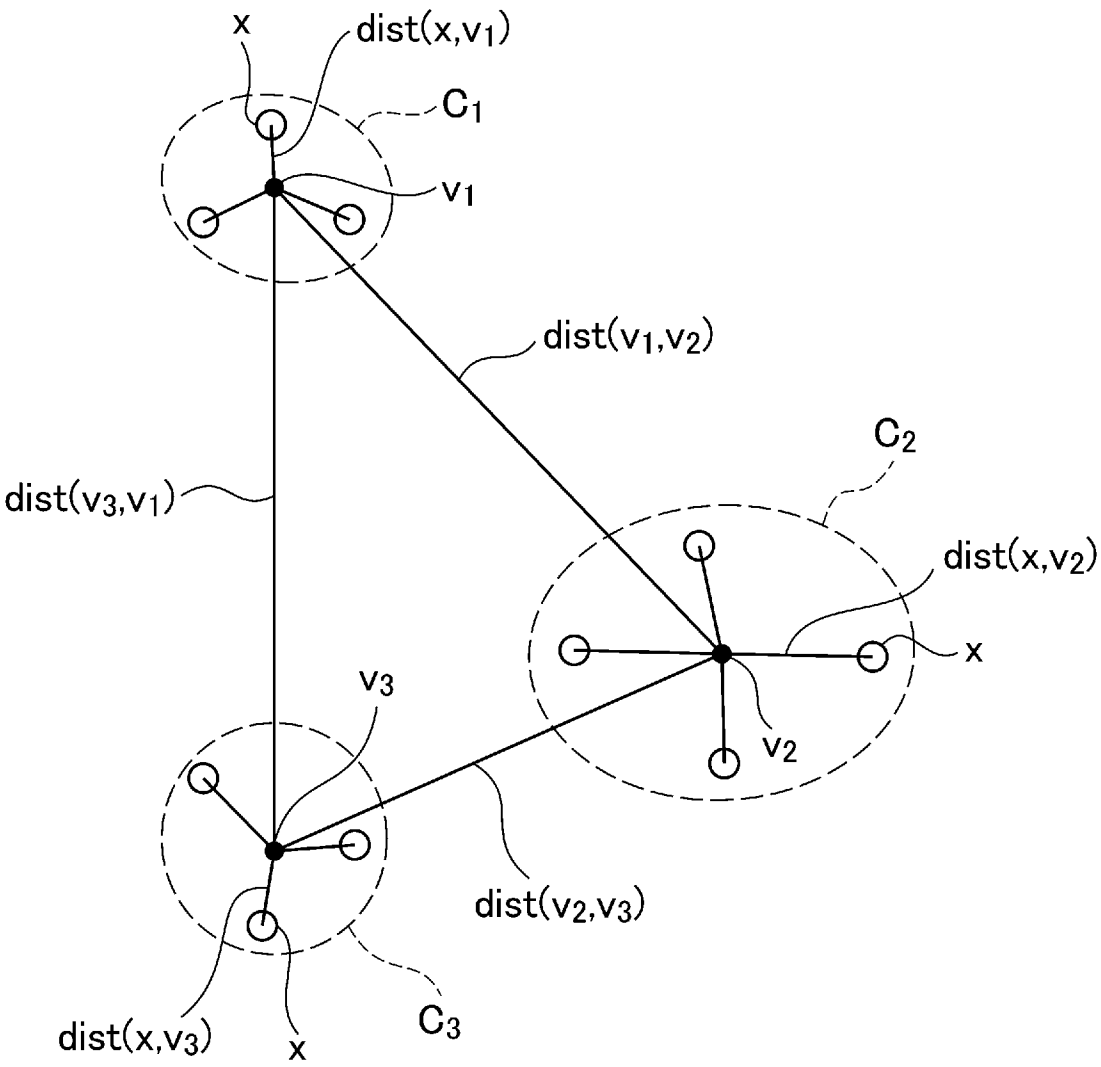
[図2]


スタート
クラスタ数cを初期化,
E0 ← 0
ST1
クラスタリング実施
ST2
評価値E1(c)を算出
ST3
評価値E1(c) ＜ E0？
NO
YES
ST4
ST5
E0 ← 評価値E1(c)
c ← c+1
ST6
クラスタ数「c−1」のクラスタリング結果を取得
（最適クラスタ数＝「c−1」）
ST7
エンド

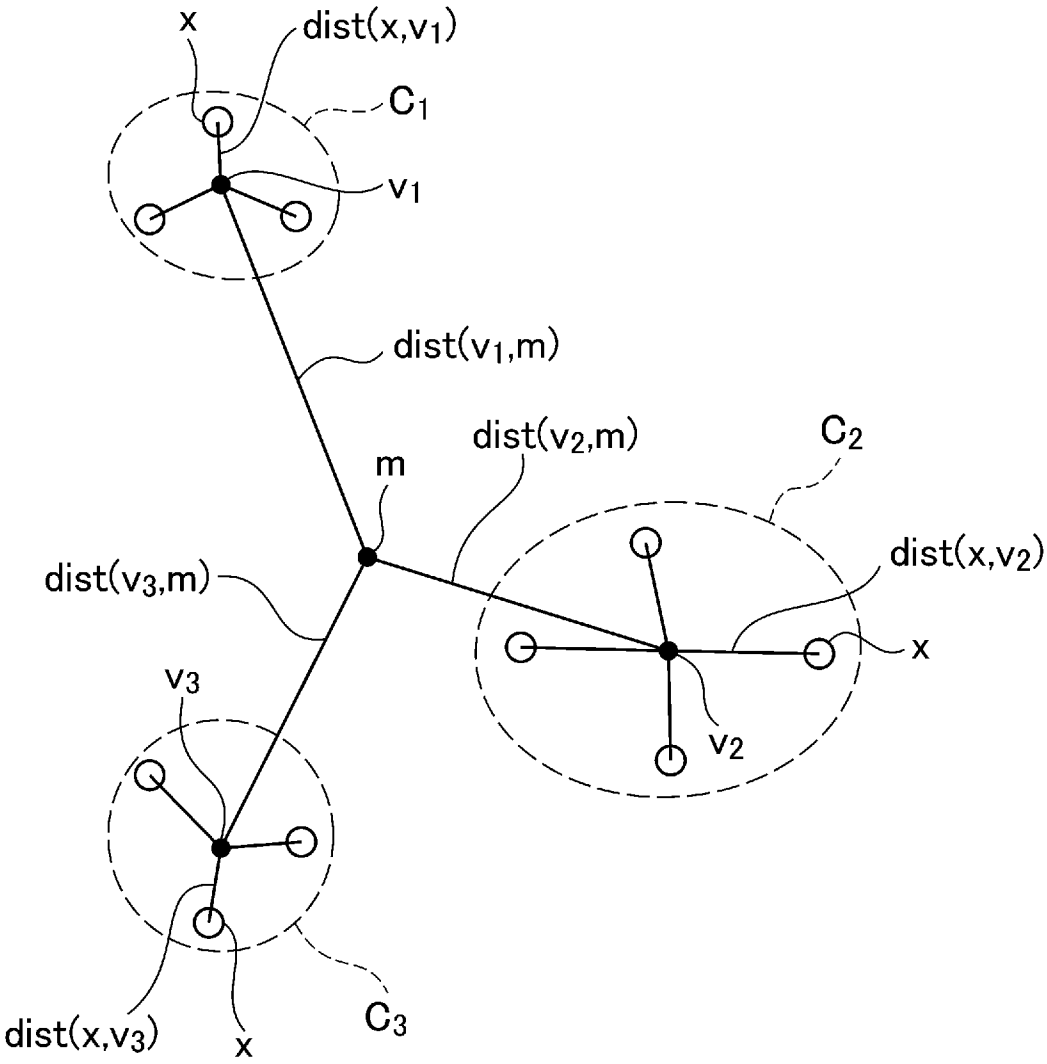
[図3]


スタート
クラスタ数cの初期値、設定最大値を設定
ST11
クラスタ数cを初期値に設定
ST12
クラスタリング実施
ST13
評価値E1(c)を算出
ST14
クラスタ数c ≧ 設定最大値？
NO
YES
ST15
ST16
c ← c+1
最大の評価値のクラスタ数の
クラスタリング結果を取得
ST17
エンド

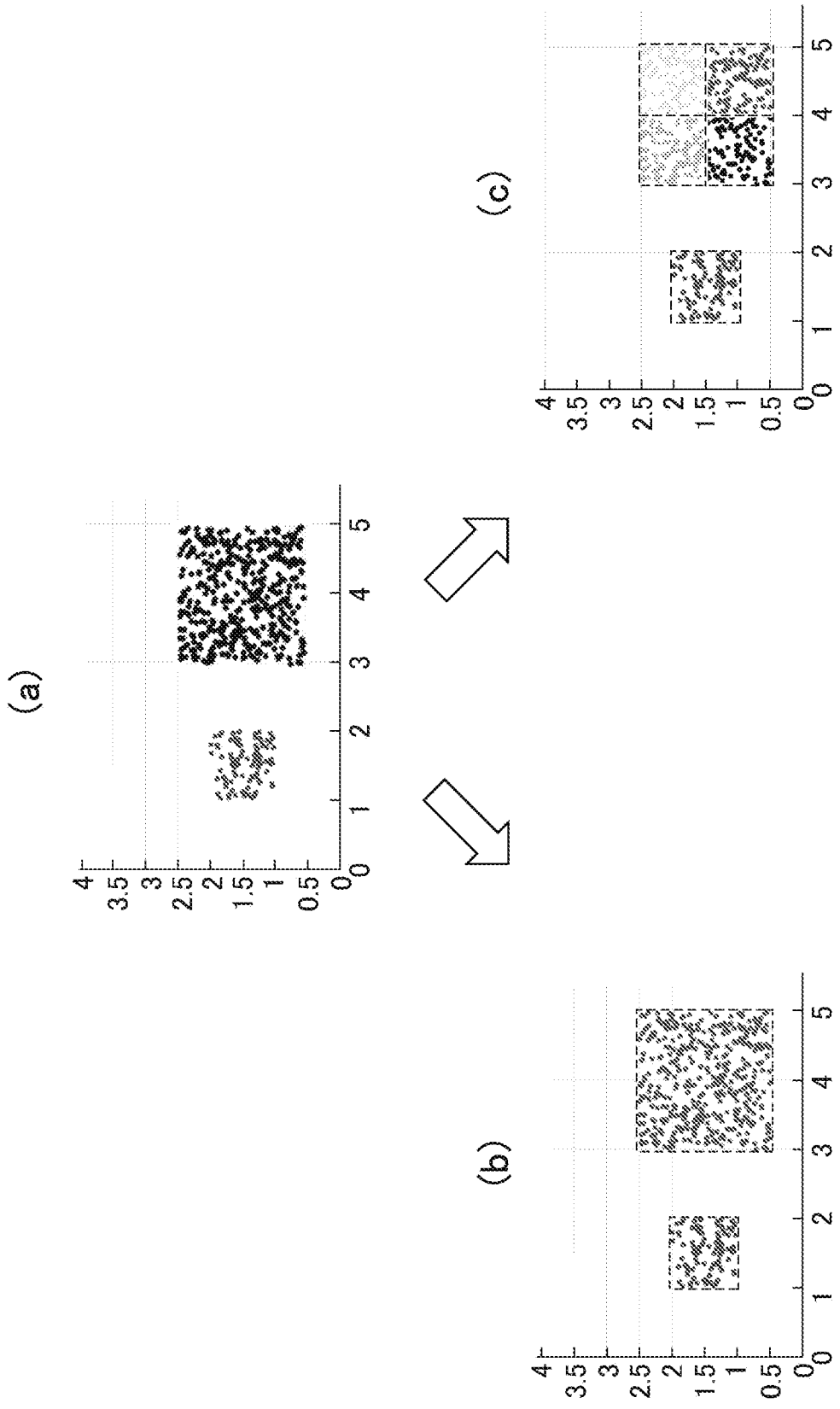
[図4]


x
dist(x,v1)
C1
v1
dist(v1,v2)
C2
dist(v3,v1)
dist(x,v2)
x
v3
v2
dist(v2,v3)
dist(x,v3)
x
C3

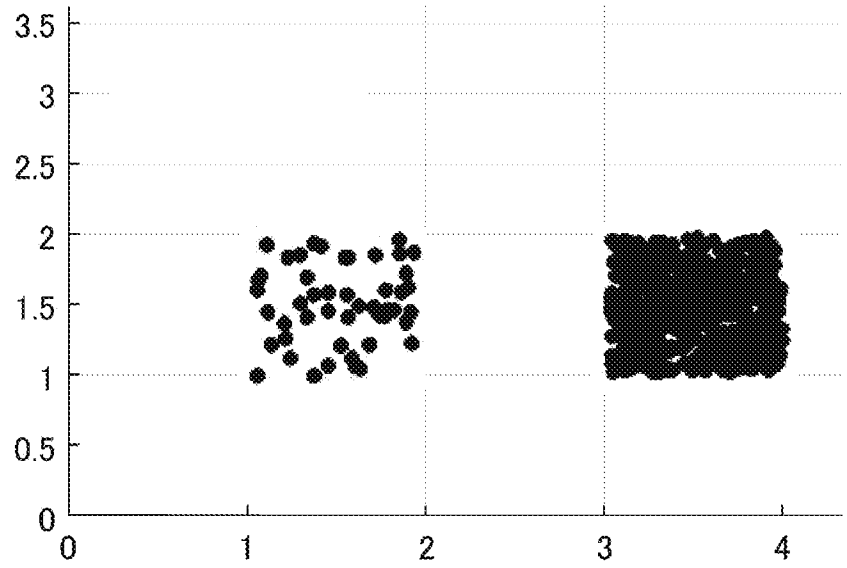
[図5]


x
dist(x,v1)
C1
v1
dist(v1,m)
dist(v2,m)
C2
m
dist(x,v2)
dist(v3,m)
x
v3
v2
dist(x,v3)
x
C3

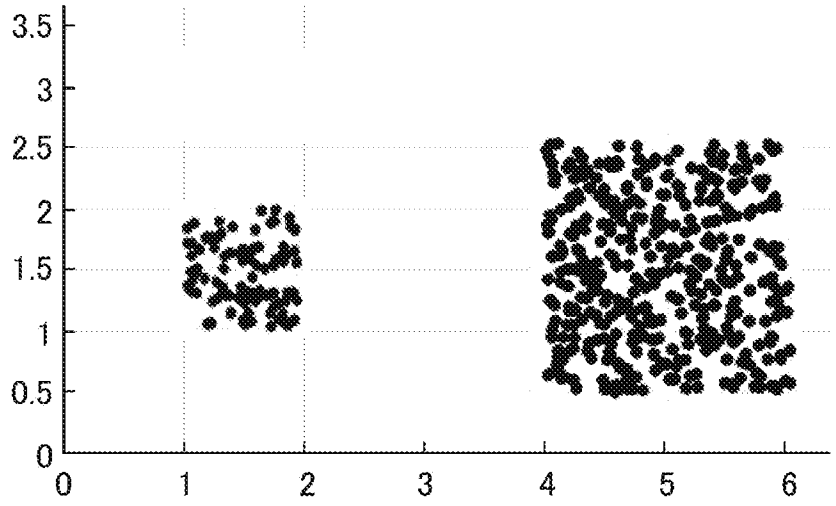
[図6]



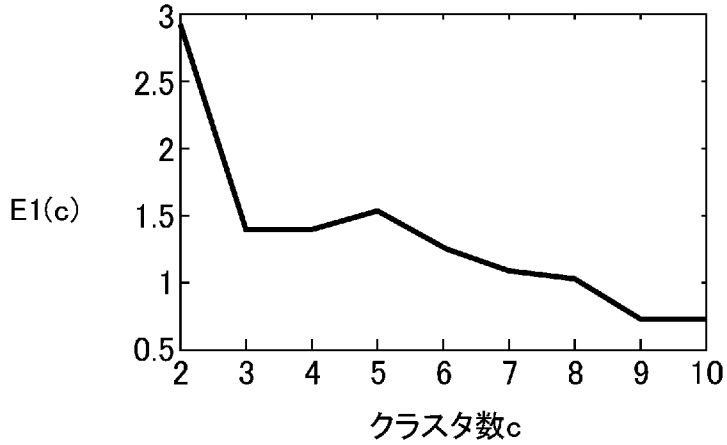
[図7]



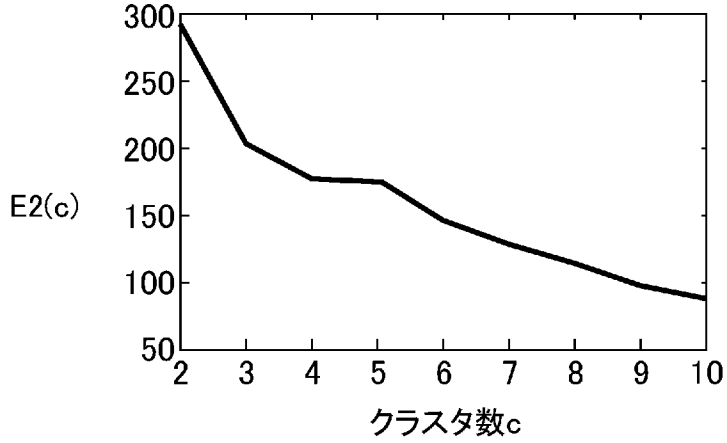
[図8]



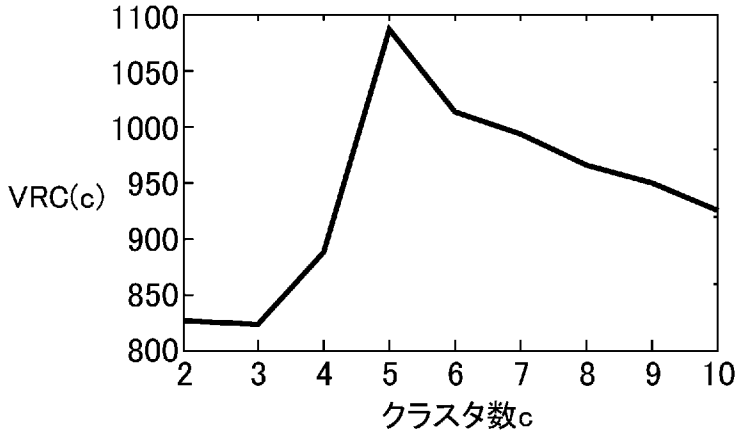
[図9]


E1(c)
3
2.5
2
1.5
1
0.5
2 3 4 5 6 7 8 9 10
クラスタ数c

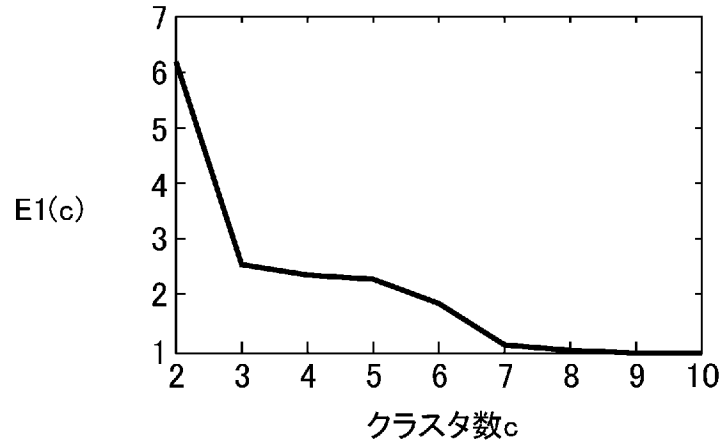
[図10]


E2(c)
300
250
200
150
100
50
2 3 4 5 6 7 8 9 10
クラスタ数c

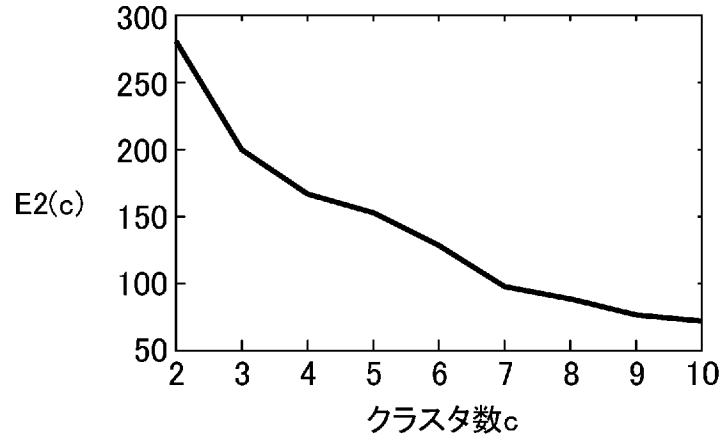
[図11]


VRC(c)
1100
1050
1000
950
900
850
800
2 3 4 5 6 7 8 9 10
クラスタ数c

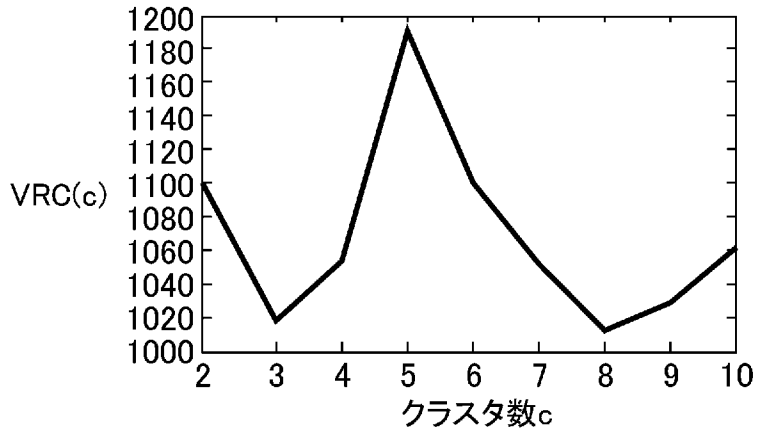
[図12]


E1(c)
7
6
5
4
3
2
1
2 3 4 5 6 7 8 9 10
クラスタ数c

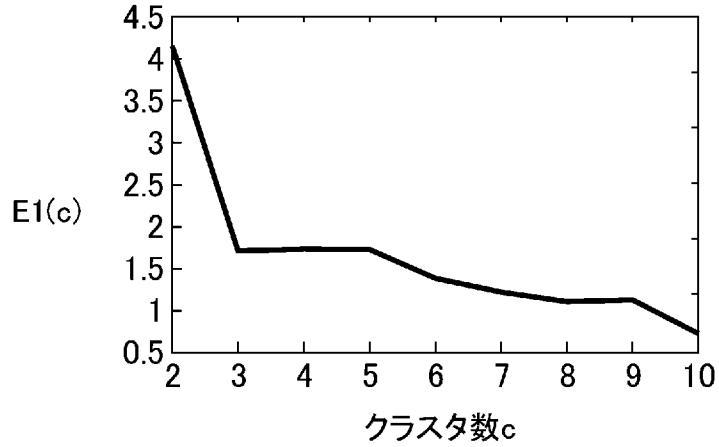
[図13]


E2(c)
300
250
200
150
100
50
2 3 4 5 6 7 8 9 10
クラスタ数c

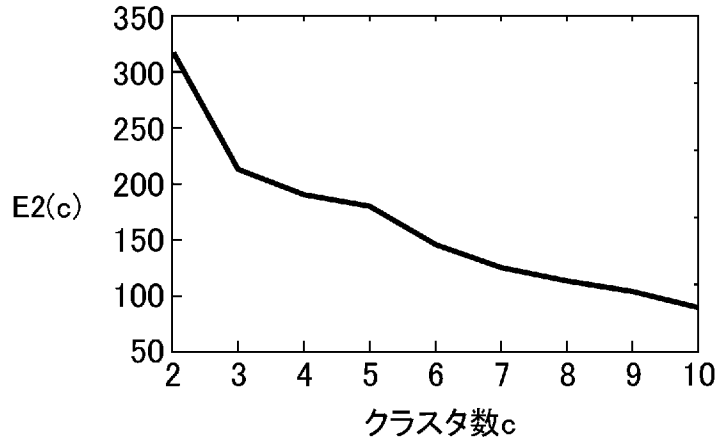
[図14]


VRC(c)
1200
1180
1160
1140
1120
1100
1080
1060
1040
1020
1000
2 3 4 5 6 7 8 9 10
クラスタ数c

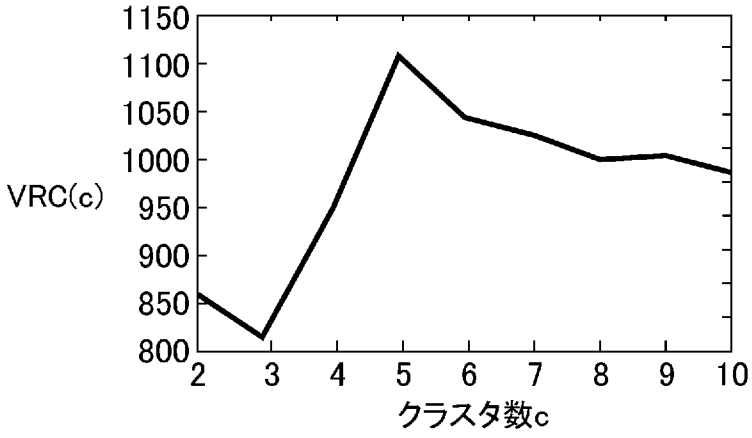
[図15]


E1(c)
4.5
4
3.5
3
2.5
2
1.5
1
0.5
2 3 4 5 6 7 8 9 10
クラスタ数c

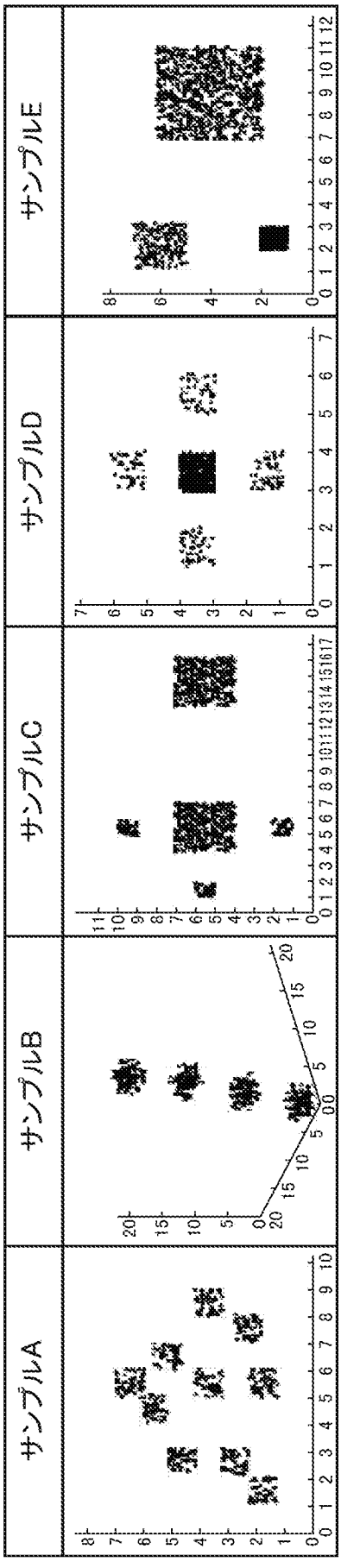
[図16]


E2(c)
350
300
250
200
150
100
50
2 3 4 5 6 7 8 9 10
クラスタ数c

[図17]


VRC(c)
1150
1100
1050
1000
950
900
850
800
2 3 4 5 6 7 8 9 10
クラスタ数c

[図18]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2018/025498

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. G06F17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl. G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan	1922-1996
Published unexamined utility model applications of Japan	1971-2018
Registered utility model specifications of Japan	1996-2018
Published registered utility model applications of Japan	1994-2018

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y A	ARBELAITZ, Olatz et al., "An extensive comparative study of cluster validity indices", Pattern Recognition, January 2013, vol. 46, Issue 1, pp. 243-256	1-4, 6-9, 11 10 5
X Y A	SAITTA, S. et al., "A comprehensive validity index for clustering", Intelligent Data Analysis, 2008, vol. 12, no. 6, pp. 529-548	1-4, 6-9, 11 10 5

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 September 2018 (12.09.2018)	25 September 2018 (25.09.2018)

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１８／０２５４９８

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１８年
日本国実用新案登録公報	１９９６－２０１８年
日本国登録実用新案公報	１９９４－２０１８年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X Y A	ARBELAITZ, Olatz et al., An extensive comparative study of cluster validity indices, Pattern Recognition, 2013.01, Vol.46, Issue 1, pp.243-256	1-4,6-9,11 10 5
X Y A	SAITTA, S. et al., A comprehensive validity index for clustering, Intelligent Data Analysis, 2008, Vol.12, No.6, pp.529-548	1-4,6-9,11 10 5

☐　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
１２．０９．２０１８	２５．０９．２０１８

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｍ　９６５２
早川　学
電話番号　０３－３５８１－１１０１　内線　３５９９