回帰分析によるクラスタリング

本吉　正博[†]　　三浦　孝夫[†]　　塩谷　　勇[††]

† 法政大学 工学研究科 電気工学専攻　〒 184–8584 東京都小金井市梶野町 3-7-2
†† 産能大学 経営情報学部　〒 259–1197 神奈川県伊勢原市上粕屋 1573
E-mail: †{i02r3243,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし　データクラスタリングは k-mean 法や各種階層的手法を初めとして様々な手法が提案されている．これらには初期値による影響や，クラスター化の基準などの問題があることが議論されている．本稿では，クラスターを結合する基準として，クラスター内の回帰分析による F 検定統計値を用いる方法を提案する．また，本手法により，局所的に異なる傾向を持つデータが分類できることを検証する．

キーワード　データマイニング, 多変量解析

Clustering by Regression Analysis

Masahiro MOTOYOSHI[†], Takao MIURA[†], and Isamu SHIOYA[††]

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
†† Department of Management and Information Science, SANNO University　1573, Kamikasuya, Isehara city, Kanagawa 259–1197 Japan
E-mail: †{i02r3243,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract　In data clustering, many approaches have been proposed. For example, K-means method and hierarchical method. A problem is in effect by initial value and criterion to combine cluster. In this investigation, we propose a method using F-value by regression analysis as criterion to combine cluster. We show that data with a local trend can be classified.

Key words　Data Mining, Multivariate analysis

1. 前　書　き

証券市場で売買される株式は，所属産業で分類されている．こうした分類は証券投資にはよく言及されることも多い．当然，それらの銘柄の値動きは似たようなものである．が，分類には時間的な安定性が弱く，産業分類の方法自体にも問題点は存在する．より明確な基準を用いて分類を行うことを分析者が意図したとき，定性的な分類を離れ，定量的な分類を試みることになる．定量的な分類を行う多変量解析の方法がクラスター分析である．

クラスター分析とは，異質なものが混じっているオブジェクトの中から，似ているものを凝集し，グループ（クラスター，cluster）分けを行うためのアルゴリズムの総称を指す．すべての研究領域で各研究者が直面している一般的な問題は，観測されたデータをいかに意味のある体系に組織立てるか，すなわち，いかに分類を行うかにある．

クラスター分析は，パターン認識（例えば地図処理での土地利用図の生成），空間データ分析，画像処理，経営分析（自動車事故のパターンから新たな保険の生成），WWW（ドキュメント分類，Weblog のクラスター化と利用パターンの抽出）等，幅広い応用分野で利用されている．

クラスター内の類似性が高く，クラスター間の類似性は低いほどよいクラスタリングである．このクラスター化の能力は，類似性の定義とその実行方法に依存する．使用された類似度が実際の類似度，あるいは，分析者にとって意味のある類似度であるかどうかについては何の保証もない．特定の応用に対して正しい方法を選択することは分析者の責任となる．また，隠れたパターンをどれだけ見出せるかが重要である．本研究では，対象として局所的に異なる傾向を持つデータを想定する．これは，局所的な部分線形空間を擁するデータ構造の推定の問題である．そしてクラスターを結合する基準として，分散とクラスター内の回帰分析による F 検定統計値を用いる方法を提案する．

次章では，既存の方法で解決できないことを論じる．第 3 章では，いくつかの定義とデータの前処理について述べる．第 4 章では，クラスターを結合する方法とその基準について述べる．第 5 章では，実験結果と他の方法との比較実験の結果を示す．第 6 章では関連する研究を簡単に紹介して第 7 章で結びとなる．

2. 局所的傾向を持つデータのクラスタリング

異なる傾向を持ったオブジェクトが混在しているようなデータのクラスタリングを考える．このようなデータは，局所的にはそれぞれが異なる線形関数で回帰でき，多次元空間では楕円状のクラスターを形成する．これらのクラスターは，もちろん交差することも考えられる．

最も簡単な解決法は，階層的手法の最近隣法を用いることにより，一番距離の近いオブジェクトを探して数珠繋ぎにクラスターを結合することができるが，クラスターが交差するような場合は，交差点でクラスターが分断されてしまう．つまり，異なる傾向を持つデータであっても，距離さえ近ければに誤って合併してしまう可能性がある．一般のミンコフスキー距離でも同様の問題を持つ．

オブジェクト集合を重心で代表する k-mean 法は本来，凸型でないクラスターには向かない．分散の基準に加え，オブジェクトが移動することで双方のクラスターの線形性が向上するような点を探したとしても，そのような点はかならずあるわけではない．クラスター数を決めなければいけないことも問題である．

２つの方法に共通した問題は，クラスター間の類似度にある．つまり，距離測度，分散の他に部分線形性を考慮した他の類似度を導入する必要がある．今までの議論をまとめると，対象データの適切なクラスタリング手法として満たすべき要件としては，

１．部分線形空間を分類可能な類似度の設定

２．解釈可能な適切なレベル（クラスター数）で収束

となる．

事前に与えられたデータに基づき未来の事象を何らかの関数により予測する多変量解析手法に回帰分析がある．クラスターに回帰分析を行い，回帰式を F 検定する．F 値が高くなるようなクラスターを選び次々に結合することで F 値が最大となるクラスターを得る．つまり本研究では，F 値をクラスター結合の類似度基準として導入する．これは，点でクラスターを代表させる手法に対して，線でクラスターを代表させる”線のクラスタリング”に相当する手法といえる．本研究では，ユークリッド距離，分散に加えこの F 値を類似度の基準とし，徐々に線形クラスターを結合しながら目的のクラスターへと’復元’していく方法をとる．次章にて，既存の統合形式の階層クラスタリングとの違いを説明する．

3. 初期クラスターの選定

’オブジェクト’とは対象世界に存在する’もの’(thing）である．我々がここで対象とするデータは，複数の変数を持つオブジェクトの集合である．変数は，すべて環境から与えられる入力であり，分類のための外的基準はないものとする（そのためデータマイニングでは，教師なし分類と呼ばれる）．変数は，２種類ある．分析者から与えられた回帰分析の際の基準となる１個の変数を基準変数（criterion variable) とし，その他の複数個の変数を説明変数（explanatory variable）とする．本研究では数値データを扱うものとするが，カテゴリカルデータについてはダミー変数や数量化理論を用いた方法が考えられる．

我々は対象データをデータ行列として扱う．オブジェクトはデータ行列の行で表現し，個々のオブジェクトが持つ説明変数，及び基準変数は列で表現する．m 個の説明変数を $x_1, x_2, \ldots, x_m$，基準変数を y とし，これらのデータが n 個得られたとする．

$$(X|Y) = \left(\begin{array}{ccc|c} x_{11} & \ldots & x_{1m} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{k1} & \ldots & x_{km} & y_k \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \ldots & x_{nm} & y_n \end{array}\right) \quad (1)$$

$(\in R^{n\times(m+1)})$

ただし，X は説明変数，Y は基準変数であり，各変数は

$$\mu_{xi} = \sum_{k=1}^{n} x_{ik} = 0 \quad ; \quad i = 1, \ldots, m \quad (2)$$

$$\mu_y = \sum_{k=1}^{n} y_k = 0 \quad (3)$$

$$\sqrt{\frac{1}{n}\sum (x_{ki} - \mu_{xi})^2} = 1 \quad ; \quad i = 1, \ldots, m \quad (4)$$

$$\sqrt{\frac{1}{n}\sum (y_k - \mu_y)^2} = 1 \quad (5)$$

となるような標準データ（z-score）に変換されているものとする．初期クラスターは，オブジェクトの集合であり，各オブジェクトは，初期クラスターに排他的に含まれる．

通常，凝集法において，最初の段階では，各オブジェクトが各クラスターを表しており，類似度はオブジェクト間の距離によって定義されるが，本手法はデータを線として扱うため，分散を持つ集合でなければならない．この最初の段階で初期クラスターを用いる．オブジェクトを小クラスターに分割し，これを初期クラスターとする．初期クラスターは内積（コサイン）により動的に求める．アルゴリズムは以下の通りである．

入力ベクトル $\vec{s_1}, \vec{s_2}, , , \vec{s_n}$ に対して

１．最初の入力オブジェクト $\vec{s_1}$ をクラスター C_1 の重心とし，$\vec{s_1}$ を C_1 のメンバとする．

２．以後，入力 $\vec{s_k}$ に対して，既存のクラスター $C_1 \ldots C_i$ との類似度を式 (6) によって計算し，どのクラスターとの類似度も閾値 $THcos$ 未満の場合は，新たなクラスターを生成してそのクラスター重心とし，類似度が $THcos$ 以上の場合は，最も類似度が高いクラスターのメンバとする．この時，メンバが新たに増減したクラスターはその重心ベクトルを式 (7) にて再計算する．

３．割り当てが終了するまで繰り返す．

4．F 値が計算できないメンバ数が m+2 個未満のクラスターは除去する．

$$cos(k,j) = \frac{\vec{s_k} \cdot \vec{c_j}}{|\vec{s_k}||\vec{c_j}|} \tag{6}$$

$$\vec{c_j} = \frac{\sum S_k \in C_j^{\vec{s_k}}}{M_j} \tag{7}$$

ここで M_j はクラスター C_j のメンバ数，m は説明変数の数である．

4. クラスターの結合

この章ではクラスター間の類似度を定義し，類似度基準が結合とどのように関係するかを述べる．我々は，ここでクラスター間の類似度を２つの側面から定義する．類似度のひとつはクラスター重心間距離である．距離測度には，回帰分析で用いる最小二乗法と同様にユークリッド距離を適用する．２または３次元空間の場合，この測度は空間内のオブジェクト間の実際の幾何的距離と一致する．

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + \ldots + |x_{im} - x_{jm}|^2 + |y_i - y_j|^2} \tag{8}$$

これによる非類似度行列は次のようになる．

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \ldots & d(n,n-1) & 0 \end{pmatrix} \tag{9}$$

$(\in R^{n \times n})$

距離が最も小さいクラスターの組み合わせは，結合するクラスターの組み合わせの候補となる．その候補が結合する上で適正かどうかを検査しなければならない．もうひとつは回帰の有効性を保つための類似度基準を回帰式の検定統計値である F 値により定義する．重回帰分析を用いた，最小二乗法による回帰式の推定と，F 検定について要約する．

F 検定は，回帰式が予測に役立つのかを F 値を用いて検定するためのものである．F 値が F 分布の数表に基づいた有意水準以上であれば，回帰式が予測に役立つといえる．(1) 式と同様のデータ行列で与えられるメンバ数 n のクラスターに対して重回帰分析のモデルは

$$y = b_1 x_1 + b_2 x_2 + \ldots + b_m x_m + e_i \tag{10}$$

である．ここで b_i の最小二乗推定量 $\tilde{b}_i$ は，

$$B = (\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_m) = (X^T X)^{-1} X^T Y \tag{11}$$

であり，これを回帰係数という．実際には，標準データを前提としているため標準化回帰係数である．y は実測値であるのに対して，回帰係数 B による予測値を Y とする．この時，回帰による変動要因について，平方和 S_R と平均平方 V_R は，

$$S_R = \sum_{k=1}^{n} (Y_k - \bar{Y})^2 \quad ; \quad V_R = \frac{S_R}{m} \tag{12}$$

残差による変動要因について，平方和 S_E と平均平方 V_E は，

$$S_E = \sum_{k=1}^{n} (y_k - Y_k)^2 \quad ; \quad V_E = \frac{S_E}{n-m-1} \tag{13}$$

この時，

$$F_0 = \frac{V_R}{V_E} \tag{14}$$

は第１自由度 m，, 第２自由度 $n-m-1$ の F 分布に従う．

次にメンバ数 a であるクラスター A とメンバ数 b であるクラスター B が得られたとき，結合クラスター $A \cup B$ を考える．データ行列は次のようになる．

$$(X|Y) = \left(\begin{array}{ccc|c} x_{A11} & \ldots & x_{A1m} & y_{A1} \\ \vdots & \ddots & \vdots & \vdots \\ x_{Aa1} & \ldots & x_{Aam} & y_{Aa} \\ x_{B11} & \ldots & x_{B1m} & y_{B1} \\ \vdots & \ddots & \vdots & \vdots \\ x_{Bb1} & \ldots & x_{Bbm} & y_{Bb} \end{array} \right) \tag{15}$$

$(\in R^{n \times (m+1)})$

ただし $n = a + b$ である．この場合も，上記と同様に式 (11) より回帰式が，式 (14) より F 値を求めることができる．結合前の２つのクラスターから得られる F 値と，結合クラスターの F 値の間の性質について下記に述べる．

クラスター A，クラスター B，結合クラスターの F 値をそれぞれ F_A，F_B，F とする．

例 1 F_A,$F_B > F$

お互いに回帰の妨げになっている場合は，傾きが有意に異なると考えられる．ゆえに，クラスター A と B の類似性は低く，クラスターの線形性は減少する．特に，$F_A = F_B$,$F = 0$ の場合は A と B の回帰式が重心を交点として直行している．

例 2 F_A,$F_B \leqq F$

２クラスターとも F 値が向上する場合は，傾きが有意に異ならないと考えられる．ゆえに，クラスター A と B の類似性は高く，クラスターの線形性は向上する．特に，$F_A = F_B$,$F = 2F_A$ の場合は，クラスター A と B のオブジェクト数と座標が全て一致している．

例 3 $F_A \leqq F$,$F_B > F$　あるいは $F_B \leqq F$,$F_A > F$

一方が向上し，一方が減少するような場合は，クラスター A と B の分散または，F 値の差が大きい時に起こる．本来結合すべきな場合も，そうでない場合も存在する．

これらより，F 値による類似度基準により結合すべきルールは，結合クラスターの F 値が結合前のクラスターの F 値に対してどちらよりも大きい場合のみであることがいえる．

ユークリッド距離を用いた非類似度は，局所的な距離以上に離れたクラスターと結合しないようにするための方法である．しかし，本手法では，F 値基準が絶対的な基準であるゆえに，

F 値基準を満足するまで距離基準を下げ続けて結合パターンを探索する．そのため，初期クラスターの不良や，もともと局所的に線形回帰できるクラスターをもたないデータの場合，距離基準をどこまでもさげてしまい，結合すべきでないクラスターが結合する可能性がでてくる．この問題を解決するために，離れてもよい距離の閾値 THd を分析者が与えるものとする．すなわち，分散基準として THd を導入する．

$THd >$（クラスター A の分散＋クラスター B の分散）×重心間の距離

F 値基準と THd を両方満足する場合，クラスターを結合できるものとする．THd はデフォルト値として初期クラスターの内分散の平均を与えるものとする．THd により，クラスターの内分散が抑えられるため，距離の遠すぎるクラスター同士が結合する心配がなくなる．

このアルゴリズムを次に示す.

(1) 全体を標準化する．

(2) $THcos$ を満たす初期クラスター計算する．メンバ数が説明変数の数 +2 未満のクラスターは捨てる．

(3) 各クラスターの重心，分散，回帰係数，F 値，全ての組み合わせについて重心間の距離を計算する．

(4) 重心間距離の近いクラスターを組み合わせの候補とし，標準化した上で回帰係数と F 値を再計算する．

(5) 結合後の F 値が結合前のクラスターのどちらよりも大きくなり，THd を満足するなら結合し，しないなら次の候補を step4 に戻って探す．もし結合できるクラスターがないならば終了．

(6) 結合したら重心と他のクラスターとの重心間距離を再計算し，step4 に戻る．

5. 実　　験

この章では，本手法の有効性を検証するため，気象データを用いた実験を行う．実験に用いたデータは，新潟気象台と稚内気象台の作成した 1997 年 1 月の気象観測時別値データを単純に連結したもので，各気象台がそれぞれ 744 件で合計 1488 件，180KB である [13]．これは，もともと局所的に線形回帰できるという初期条件の元で本手法を適用するためである．

データは，1 時間ごとに観測された合計 22 項目のデータで，この実験ではそのなかから数値データでしかも欠損値のない，日（day），時（hour），現地気圧（hPa），海面気圧（hPa），気温（℃），露点温度（℃），蒸気圧（hPa），相対湿度（%）の 8 項目を変数の対象とする．この他に観測地点番号があり，クラスタリングの評価にのみ用いる．幾つかのデータを表 1 に示す.

実験では，すべての変数を標準化し，提案するアルゴリズムによって解析した．基準変数は，気温としその他の変数は説明変数とした．初期値決定のための閾値 $THcos$ は 0.8，クラスター間距離の閾値 THd は 15 とした．結果を表 2 に示す．コサインによる初期値決定で，1364 個のオブジェクトから 40 個の初期クラスターが得られた．残り 124 個のオブジェクトは，十分なメンバ数を持つクラスターに分類されなかったため除外された．結合ループは 35 回で収束し，5 個のクラスターを得た．

表 1　気象データ

地点	日	時	現地気圧	海面気圧	気温	...
604	1	1	1019.2	1020	5	...
604	1	2	1018.6	1019.4	5.2	...
604	1	3	1018.3	1019.1	5.4	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮
401	31	24	1014.6	1016	-5.8	...

クラスター 1 では，19 個の初期クラスターが結合された．クラスター 2 と 3 は結合されなかったクラスターである．クラスター 4 では，10 個の初期クラスターが結合された．クラスター 5 では，9 個のクラスターが結合された．一般的に言って，この結果は観測地点による特徴を示している．実際，クラスター 1 は新潟地点のオブジェクト 744 個のうち 519 個のオブジェクトを含んでいる．クラスター 5 は，稚内地点のオブジェクト 744 件のうち 469 件を含んでいる．このことから，クラスター 1 は新潟地点，クラスター 5 は稚内地点固有の傾向を持つオブジェクトをそれぞれよく抽出していることがわかる．

例えば表 3 で示される各クラスターの重心から，クラスター 1 は他のクラスターと比較して気圧と気温が高い．ゆえに，高度が高く，しかも気温の高い地域で観測されたと推測できる．また，クラスター 5 は湿度と気温が低い．ゆえに，降水量が低いか乾燥地帯であり，しかも気温が低い地域で観測されたと推測できる．クラスター 4 は，日が高い，つまり 1 月後半であり，気圧が低く，湿度が高い．ゆえに，このクラスターは観測地域とは関係ない，低気圧などの気象状態を特徴とするクラスターであると推測できる．実際，クラスター 4 は新潟地点と稚内地点のオブジェクトをほぼ同数持っている．

また，表 4 で示される各クラスターの回帰係数より，クラスター 5 は 1 と 4 に比べて全体的に係数の絶対値が高いことから，気象状態の変化に対して，気温変化が大きい．つまり，気温の取りうる幅が広いクラスターであるといえる．北海道は，冬期に日最低気温が-20 ℃近くまで低下し，夏期に日最高気温が 30 ℃を超えるなど年較差が最大の地域である．この結果は，実際の分類に当てはまっているといえる．クラスター 1 はクラスター 5 と傾向が似ているが，より傾きが小さい．気温の取りうる幅の狭いクラスターであるといえる．クラスター 4 は，他のクラスターと異なり，露天温度と相対湿度のみ相関がある．

本実験をまとめる．最終クラスターは 5 つ得られた．特に，クラスター 1 と 5 から地域的な特徴を持つ傾向を抽出することができた．これによる知見から，クラスタリングにより，オブジェクトが適切に分類されたかどうかは，表 2 の観測地点の情報から明らかである．ゆえに，本実験によりデータが初期条件を満足することを示せた．

表 2　最終クラスター（$THcos = 0.8, THd = 15$）

	内分散	F 値	所属クラスター数	新潟	稚内
Cluster1	4.958	8613.28	19	519	74
Cluster2	2.12926	2043.62	1	29	1
Cluster3	2.42196	78.2235	1	45	0
Cluster4	5.1034	85603.6	10	135	189
Cluster5	5.50085	17964.9	9	11	469

表 3　クラスター重心

	Cluster1	Cluster4	Cluster5
日	-0.0103696	0.487242	-0.0987597
時	0.0622433	-0.148476	0.0358148
現地気圧	0.599712	-0.899464	-0.0542843
海面気圧	0.580779	-0.902211	-0.023735
露天温度	0.512113	0.294745	-1.05024
蒸気圧	0.468126	0.245389	-0.993099
相対湿度	-0.179054	0.975926	-0.392923
気温	0.692525	-0.234597	-0.976859

表 4　クラスターの標準化回帰係数

	Cluster1	Cluster4	Cluster5
日	-0.0163907	-0.0029574	0.0108929
時	-0.00316974	-0.00182585	-0.00965708
現地気圧	1.18154	0.0357092	1.77679
海面気圧	-1.15683	-0.0393822	-1.77494
露天温度	0.909799	1.103	1.22524
蒸気圧	0.421526	-0.016361	0.212142
相対湿度	-1.25678	-0.36817	-0.804252

比較のために，他の手法で同じデータを解析した．解析は統計解析アプリケーション SPSS の k-mean 法を用いた．初期クラスター中心は乱数で決定される．最大反復回数は 10 回に設定した．k が 2 個，3 個の場合についてそれぞれ解析を行った．k=2 の結果を表 5，6，k=3 の結果を表 7，8 に示す．k が 2 個の場合，得られた 2 つのクラスターを観測地点に関しての有意な違いをこの結果から見出すことは困難である．

表 5　K-mean 法による分類（k=2）

	新潟	稚内
Cluster1	496	359
Cluster2	248	385

表 6　クラスターの重心（k=2）

	Cluster1	Cluster2
日	13	21
時	12.8	12.1
現地気圧	1016.98	1001.69
海面気圧	1018.03	1002.86
露天温度	-4.29	-4.38
蒸気圧	4.68	4.62
相対湿度	69.0	73.3
気温	0.95	0.79

k が 3 個の場合も同様である．従って，本手法は K-mean 法でうまくクラスタリングできない場合の代替手段となりえる．

表 7　K-mean 法による分類（k=3）

	新潟	稚内
Cluster1	293	149
Cluster2	158	353
Cluster3	293	243

表 8　クラスター重心（k=3）

	Cluster1	Cluster2	Cluster3
日	12	21	14
時	12.5	12.0	14
現地気圧	1017.50	1000.20	1014.49
海面気圧	1018.50	1001.42	1015.56
露天温度	-1.41	-4.71	-6.37
蒸気圧	5.66	4.49	3.98
相対湿度	81.1	78.4	59.8
気温	1.52	-1.36	0.64

6. 関 連 研 究

クラスター分析の分類手法は大きく非階層的手法と階層的手法とに分かれている．非階層的手法は評価尺度に最も適合するような，k 個のクラスターに分割する方法である．この大域的最適解はすべてのパターンを探索するため現実的でない．発見的な方法としてクラスターを中心で表す k-mean 方法 [1] や事例のファジィ分類が可能な fuzzy k-mean 法 [2]，クラスター数を自動的に決定可能で事例の確率的な分類を行う AutoClass [3] や Snob [4] などが知られている．

k-mean 法は比較的速く，計算量は，$O(tkn)$ である．ただし，n:オブジェクト数，k:クラスター数，t:繰り返し数で通常 k，$t \ll n$ である．欠点は予め K が決められている，結果が初期値に依存する，局所解の可能性がある，雑音，誤りに無防備，凸型でないデータに適用できないなどが挙げられる．一方 SOM [5] では，予めクラスターに低次元の近傍関係を与えておくことによりクラスター同士の遠近を低次元表現で示すことを可能にしている．

階層方式には，統合方式（Agglomerative Nesting）と分割方式（Divisive Analysis）がある．統合方式は，凝集法（ツリークラスタリング）とも呼ばれる．統合方式は，距離による非類似度行列を用いて類似するノードを次々と合併する方法で，これを Single-Link 方式という．問題は，どのレベルのクラスターを選ぶかである．欠点は，計算量 $O(n^2)$ であること，一度確定した統合を見直せないことである．距離方式との併用による拡張には，部分クラスターの質を次第に向上させる BIRCH [9] やクラスターの代表元を抽出し，それに向かってクラスターを収縮させる CURE [10]，動的モデリングを採用する CHAMELEON [8] がある．分割方式は統合方式の逆順序で，最終的には単一要素集合になる．

逐次的にクラスターを構築する手法は ” 概念形成 ” と呼ばれ代表的なものとしては COBWEB [6] や，CLASSIT [7] などが

知られている．

また多変量解析をクラスタリングに適用する方法として，主成分分析を用いて局所的に次元を縮小する LDR [12] がある．

7. 結　　論

本稿では，異なる局所的な傾向を持つオブジェクトが混在したデータに対して，有効なデータクラスタリングの新たな手法を提案した．クラスターの傾向を回帰分析を用いて抽出し，回帰式の F 値をクラスターの類似度として定義した．クラスターの精度を保つためのクラスター間距離の閾値を定義した．また，実験により，解釈する上で現実に即した数のクラスターとその特徴を重心と回帰係数により抽出した．他の手法との比較実験を行い，この手法の有効性を示した．

筆者らは、既に時系列データから時区間を時制クラスとして抽出する技術を論じている [11]．今後は、本稿で示した手法との融合を図り、時系列データ、ストリームデータに統合化する方法を展開する予定である．

謝　　辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 14580392) の支援による．

文　　献

[1] MacQueen, J.B.: "Some methods for classification and analysis of multivariate observations" , proc. *Fifth Berkeley Symposium observations, ProStatistics and Probability*,1,University of California Press,1967.

[2] Bezdek,J.C.: "Numerical taxonomy with fuzzy sets", *Journal of Mathematical Biology*,1,pp.57-71, 1974.

[3] Cheeseman,P.,et al: "Bayesian classification", proc. *American Association of Artificial Intelligence 7th. Annual Conf.on A.I.*., pp.607-611,1988.

[4] Wallace,C.S.and Dowe, D.L.: "Intrinsic classification by MML-the Snob program", proc. *7th Australian Joint Conference on Artificial Intelligence*,pp.37-44,1994.

[5] Kohonen,T.: "Self-Organization and Associative Memory", *Springer-Verlag* New York,1989.

[6] Fissher,D.H.: "Knowledge acquisition via incremental conceptual clustering", *Machine Learning* 2, pp.139-172,1987.

[7] Gennari,J.H.,Langley,P. and Fisher,D.: "Models of incremental concept formation", *Artificial Intelligence*,Vol.40,No.1-3,pp.11-62, 1989.

[8] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: "Hierachical clustering using dynamic modeling", *Computer*, 32(8):68-75, August 1999.

[9] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: "An efficient data clustering method for very large databases", *SIGMOD Record*,25(2):103-114, June 1996. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data.

[10] S. Guha, R. Rastogi, and K. Shim. CURE: "An efficient clustering algorithm for large databases", proc. *of ACM SIGMOD International Conference on Management of Data*, volume 27, pages 73-84, New York, 1998. ACM, ACM Press.

[11] Motoyoshi,M., Miura,T., Watanabe,K., Shioya,I.: "Mining Temporal Classes from Time Series Data", proc.ACM *Conf. on Information and Knowledge Management* (CIKM), 2002.

[12] Chakrabarti,K. and Mehrotra,S.: "Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces", VLDB 2000, pp.89-100, 2000.

[13] 日本気象協会編：気象データひまわり，丸善，1998