水文データへのパターン分類化手法の適用性評価

小尻利治*・T.E. ウニー**

Application of Classification Techniques to Hydrological Data

T. Kojiri, T.E. Unny

To handle the time-dependent and multi-point properties of hydrological data, all data should be treated and analyzed in vector. Pattern recognition theory is powerful and useful for these concepts. For classification, there are several important circumstances, such as (i) the objective functions suited to hydrological analysis, (ii) performance of the global optimum and (iii) optimum number of clusters. In this paper, random K-means algorithm is introduced giving many different initial cluster centers and compared with DHB and ABF algorithms. On the other hand, three objective functions are proposed. The first is to estimate the whole shape as the distance between sample data and cluster center in time. The second is to estimate the peak value having great influence on flood planning. The third is to estimate the gradient of sequence. Then the aggregate objective function is formulated by maximizing three objective functions. Finally, the optimum number of clusters is decided under the condition of constraints subjected to the intra distance represented by aggregate objective function and the inter distance. These threshold values are gained by comparing multi-optimization and AIC techniques.

Key words: time sequence of hydrological data, pattern classification, random K-means, multi-optimization, AIC

降水、流量などの水文事象を時系列としてとらえると、ベクトルパターンとして表現することができる。Unny や宇治橋らは、パターン認識概要を導入して、降雨、流量系列の特性把握を試みてきた。しかし、分類にあたっては、(i)解析に適した目的関数、(ii)グローバルな最適解を達成する手法、(iii)適切なクラスター数の決定が必要である。そこで本研究では、分類化手法の一種である K-means を用い、解の改善策として、random K-means と DHB, ABF による分類過程を比較・検討した。また、目的関数は、時間的相違の最大値、ピーク値の相違、変化率の相違をベースとして、それらの最大値とした。この目的関数を最小化するように分類が達成される。クラスター数に関しては、内部距離(目的関数)と外部距離による多目的最適化の利用、内部距離が正規分布するとして AIC で判定する方法、および、制約による方法を提案した。最後に、実流況での適用を通じて、水文事象の分類化手順を明らかにした。

キーワード：水文時系列、パターン分類、ランダム K-means、多目的最適化、情報量基準 (AIC)

I. 序 論

水文事象を一つのベクトルパターンとしてとらえ、パターン形状の相違によってその水文特性を評価しようとする試みは、Unny ら (Unny et al, 1981)

によって提案された。その後、宇治橋ら (宇治橋 他, 1988) は、流量だけでなく降雨系列の分類へと適用範囲を広げている。しかし、パターン分類は、同じデータを扱うにしても解析時の目的関数によって結果が異なり、水文事象での分類目的に適した関数形の設定が必要である。また、従来の分類化手法では、繰り返し計算のため必ずしも全体的な最適解に到達する保証はなく、局所解となる場合が多い。そこで本研究では、Ismail ら (Ismail et al, 1988) によって提案された random K-means 手法を導入してそ

*岐阜大学工学部

**ウオタール大学工学部

*Univ. of Gifu

**Univ. of Waterloo

の収束特性を考察するとともに、いくつかの目的関数を提案し、水文学的観点から見たパターン認識概念の可能性をさぐるものである。さらに、サンプルデータの分布を考慮した、最適なクラスター数の決め方についても考察を加える。

II. パターン分類化手法の概要

パターン分類の演算方法の一つに K-means 手法がある (Tou et al, 1974)。この方法は各クラスターのセンターをそのクラスターに属するサンプルの平均値で代表し、クラスターセンターとサンプル間の距離 (目的関数) を最小化するものである。具体的には、まず、時系列での水文データをベクトルで表わし、サンプルデータと名付ける。これに対し、分類された 1 つの集合をクラスターといい、その代表ベクトルがクラスターセンターである。K-means 手法では、クラスター内のサンプルデータを時間別に平均化してセンターを求めている。一般に目的関数は、Euclidean 距離の二乗和が用いられている。サンプルがどのクラスターに属するかは、センターとの距離だけで決まっている。さらに、その分類されたクラスターの中で平均化を行ない新しいセンターを決めていくわけである。ところが、初期センターのとり方によっては局所解に収束し、目的関数に合致した分類がなされない場合がある。random K-means では、初期クラスターセンターはサンプルデータの中より一様乱数で与えられる。以後は、従来の K-means と同様にして収束するまで繰り返しセンターが求められる。一つのパターン分類が終了すると、続いて、新たな乱数によって別の初期センターを与え、再度、解を求める。こうした手順を多数回繰り返し、全解集合のうち解析目的にあったもの、すなわち、最小の目的関数値をとる場合が最適解となる。

DHB (Ismail et al, 1988) は、random K-means におけるクラスターセンター改善法の 1 種であり、サンプルデータの新しい属性を次のようにして与える。すなわち、いま K 個のクラスターに分類している時、クラスター s に属するサンプル $X(i)$ が、次の条件が満足するならば、クラスター r に移行させるものである。すなわち、

$$\frac{V_s |X(i) - Z(s)|^2}{V_s - 1} > \frac{V_r |X(i) - Z(r)|^2}{V_r + 1} \quad (1)$$

$$\frac{V_r |X(i) - Z(r)|^2}{V_r + 1} = \min \left\{ \frac{V_j |X(i) - Z(r)|^2}{V_j + 1} \right\} \quad 1 < j < K, j \neq s \quad (2)$$

である。ここに、 V_j はクラスター j でのサンプル数である。加えて、式(2)で最小化をとらないで、式(1)の状態が発生したら、ただちに、属するクラスターを r に変更する ABF も同時に提案されている。 $X(i)$ と $Z(j)$ は、それぞれ、時刻 t 、サンプル i とクラスターセンター j の要素 $X(i, t)$ (たとえば、流量や降水量) と $z(j, t)$ から構成されるベクトルである。

III. 目的関数

水文時系列間の相違を抽出するために、次の 3 種類の目的関数を提案しよう。最初の目的関数は、距離として系列全体の相違を表現するものである。二つの形状 (例えば、ハイドログラフ) のずれのうちで、時間軸上の最大値で与えられ、次のように定式化される。

$$OF_1(X(i), Z(j)) = \max_t \left\{ \frac{|x(i, t) - z(j, t)|}{z(j, t)} \right\} \quad (3)$$

なお、距離はセンターの要素によって正規化される。

第二の目的関数は系列のピーク値で比較するものである。なぜなら、ピーク流量は治水計画の策定や貯水池操作のルールベース化に際しては、重要な指標となるからである。具体的には、

$$OF_2(X(i), Z(j)) = \frac{|xp(i) - zp(j)|}{zp(j)} \quad (4)$$

となる。ここに、 $xp(i)$ はサンプル i のピーク値である。第三の目的関数は、各時刻における系列の変化率を評価するものである。目的関数(1)で同じ値であっても、変動が異なれば別の特性を有すると判断できるよう。やはり、センターの変化量によって正規化すると、目的関数は次のようになる。

$$OF_3(X(i), Z(j)) = \max_t \left\{ \frac{|(x(i, t) - x(i, t+1))|}{|z(j, t) - z(j, t+1)|} \right\} \quad (5)$$

最後に、3 種類の目的関数は次のように統合化される。

$$\begin{aligned} OF_a(X(i), Z(j)) \\ = \max \{ w_1 OF_1(X(i), Z(j)), \\ w_2 OF_2(X(i), Z(j)), w_3 OF_3(X(i), Z(j)) \} \end{aligned}$$

(6)

ここで、 w は上述の三つの関数を同じスケールで比較するための重みである。経験より、 $w_1=w_2=1.0$ 、 $w_3=0.1$ を用いている。式(6)によって、異なった季節、異なった現象、異なった観測点でのデータを同時に評価することができる。なお、目的関数 OF_a は非線形性が著しいので、次の二種類の関数でも分類を行い、比較・考察を加えておく。

$$OF_n(X(i), Z(j, u)) = \sum_{t=1}^T \frac{\{x(i, t) - Z(j, t)\}^2}{z(j, t)} \quad (7)$$

$$OF_s(X(i), Z(j, u)) = \sum_{t=1}^T \{(i, t) - z(j, t)\}^2 \quad (8)$$

両式とも二乗和型で表現されているが、式(7)はクラスターセンターで正規化され、式(8)は非線形性が少ない *Euclidean* 距離である。

IV. 最適クラスター数の決定

1. 制 約 法

水文データが任意のクラスター数（たとえば K 個）に分類されると各クラスターでの目的関数値は分類された集合の広がりを表わすことになる。ここで、 K 個のクラスターでの目的関数の最大値を内部距離 $DK(K)$ と再定義しよう。同時に、クラスター間の距離を外部距離 $EK(K)$ とする。数学的には、クラスターセンターによって次のように定式化される。

$$DK(K) = \max_{j, i \in j} \{OF_a(X(i), Z(j))\} \quad (9)$$

$$EK(K) = \min_{j, j', j \neq j'} \{OF_a(Z(j), Z(j'))\} \quad (10)$$

全ての計算可能なクラスター数(K)に対して内部距離、外部距離が計算されると、水文事象に適したクラスター数を決めなければならない。通常は、分類すべき数が与えられ、内部距離、外部距離はそれらに課せられた制約を満たせば良い。例えば、

- (i) 分類された各集合の広がり（内部距離）は3以下、
- (ii) 集合間の相違（外部距離）は1以上、
- (iii) クラスター数は全サンプルデータの半数以下である。本研究では、上記のように与えられた条件を満たす分類のうちで、最小のクラスター数を最適とする方法を提案し、制約法と呼ぶことにする。但し、その制約値には明確な基準はなく、次節の決定方法を参考に与えられる。

2. 多目的最適化手法の適用

いま、パターン分類の目的を少し緩和し、内部距離に関しては最小化、外部距離に関しては最大化と定義しよう。具体的な目的関数は、

$$\begin{cases} DK(K) \rightarrow \min \\ EK(K) \rightarrow \max \end{cases} \quad (11)$$

とベクトルで表現される。 $DK-EK$ 平面上に ($DK(K), EK(K)$) ($K=1, 2, \dots, KK$; KK は分類されたクラスター数の最大値) の点をプロットしていくと、多目的計画法でいう *Transformation Curve (TC)* が得られる。ここで、目標 (*Goal*) より *Indifference Curve (IC)* を描き、 TC と接する点が最適解になる。同時に、対応するクラスター数が最適数である。接点を得るためには、次式より EK 座標を変換すれば、*Goal* を中心とする同心円が IC となる (Kojiri et al, 1986)。すなわち、

$$\gamma = \sqrt{\frac{(e1-e0)^2 - (e1-e0)^3}{(d1-d0)^2 - (d2-d0)^2}} \quad (12)$$

$$\text{when } e2 > e1 \text{ and } d1 > d2$$

$$\text{or } e2 < e1 \text{ and } d1 < d2$$

である。ただし、($d0, e1$) は目標の座標、($d1, e1$)、($d2, e2$) は、それぞれ、 TC の両端点を示す。目標値は、内部距離に関しては、最も短い場合、すなわち、全てのサンプルデータがクラスターセンターとなった内部距離ゼロとすることができる。一方、一つのクラスターで全サンプルが包含できると、最小数での分類が最も効率的であることを意味している。すると、サンプルに隣接する未知データとの外部距離は、そのクラスターの内部距離に等しいとみなせるので、外部距離の目標値はクラスター数1の場合の内部距離で表わせる。

3. 情報量基準の導入

サンプルデータの目的関数値がクラスターセンターのまわりで正規分布をしていると仮定すると、クラスター j での対数尤度関数 $W(j)$ は

$$W(j) = -\frac{1}{2} \sum_{i=1}^{V_j} \{\ln(2\pi) + OF_a(X(i), Z(j))\}^2 \quad (13)$$

となる。ここに、分布は正規化されているので、平均0、標準偏差1とする。全情報量は、 $W(j)$ の和で与えられるので、最適クラスター数は、次の最小化問題となる。

$$AIC = -2 \sum_{j=1}^K W(j) + 2K \rightarrow \min \quad (14)$$

ここに、第2項の K は自由度を意味し、クラスター数が相当する。

以上のように、制約法では、与えるべき制約値が明白でなく、多目的最適化では、全ての可能なクラスター数で評価しなければならない問題点がある。これに対して、*AIC* では、任意のクラスター数毎に評価値が得られ、情報量の最小値が得られたと判断できれば計算を打ち切ることができ、実用上便利である。

V. 実流況での適用と考察

1. パターン分類化手順

カナダ、オンタリオ州南西部に位置する Thames 川流域の流量データ (Thamesville 地点) を用いて分類手法の効率と目的関数の適合性を評価しよう (図-1 参照, 観測点流域 4300km²). 図-2 は 1952 年 10 月から 1967 年 9 月までの 15 年間の観測データを示し, 数値はデータ番号である. 10 月から 3 月までの乾期で, 月流量は対数正規分布に従うとみなし, 対数変換したデータで適用を行なった. また, センター数は 4 個と固定し, 50 個の異なる初期センターを発生させた. 図-3 は目的関数 OF_a を random K-means によって求めた結果である. 図中上部の値は, 異常に大きい目的関数値を表わし, Run number

图—1 Thamas 川流域

図-2 サンプルデータ

につれて右上がりに書かれている。また、図中下部の値は、系列中の最小値である。OFa では、ほぼ2回の最適解が得られているが、全体に激しく変動し

図-3 random K-means による OFa の変化

図-4 OF₁における random K-means と DHB の比較

図-5 OF_nにおける random K-means との DHB の比較

ている。従って、以下では、演算の効率を見るため、非線形性を1つ減らした OF_1 を用いて議論を進める。図-4は目的関数 OF_1 を使い、 DHB と random K-means での比較である。 DHB は変動がはげしく、初期センターによっては、最適解が得られない場合が多いことを示している。 ABF も同様な変動を示している。random K-means は他の方法よりも変動が小さいので、最小化型の目的関数には適合するといえる。図-5は、目的関数 OF_n を使い、random K-means と DHB で計算した結果である。こんどは、 DHB の方が変動の少ない結果となっており、正規化を含む合計型の目的関数に適している。図-6は目的関数 OF_n と OF_s を用いた ABF の結果である。 OF_s の方が変動が小さく、非線形性の少ない評価関数には OF_s は極めて有効になる。

表-1は分類されたセンターのサンプル番号である。もとの流量系列を見ると、データ2, 4, 7, 12, 13, 14, 15には時刻 $4 \leq t \leq 5$ において増加の傾向が見られる。データ3, 6, 8, 10, 11には同時刻に減少の傾向が見られる。データ6は全体的にデータ8に類似している。データ4もまた、総流出量が異なっているにもかかわらず、形状は観測データ11に類似している。形状から考えると、データ3, 6, 8は同じクラスターに属するはずである。データ4, 11はまた別のクラスターに属するものである。しかしながら、データ7, 12, 13はすべての時刻において単調増加している。結局、総合的な目的関数 OF_d を用いれば、こうした水文学的観点を満たす分類が達成されることがわかる。ただし、この結果は、表-2に示される相互相関分析の結果とは異なっている。すなわち、

図-6 ABF での OF_n と OF_s の比較

相関係数が高くても、形状全体から見ると別のグループと見なされるわけである。この点がパターン認識概念の特徴であり、有効性と言えよう。

2. クラスタ数決定

図-7は3種類の評価値（内部距離、外部距離、AIC）の関係を示したものである。内部距離がほぼ単調に減少しているので、各クラスター数で最適な分類ができたと言える。しかし、外部距離が変動しているのは、内部距離だけが目的関数になっており、同じ内部距離を有するクラスターセンターの組み合わせに対して、外部距離を最大にする組み合わせが選ばれてないからであろう。AICは滑らかな曲線を示しており、クラスター数4で最小値をとっている。図-8は、多目的最適化でのTCとICである。ただし、一方の目的が同じ値の時に、他の目的が異なる数値をとる場合は、パレート最適解（TC上の点）となる最善の組み合わせだけを選んでプロットしている。適用例では、式(12)中の条件が満たされていないので、ICは両端点を通る直線に平行に描かれる。最適値は、AICとおなじクラスター数4であった。これらの結果より、制約法での条件は、内部距離に関しては、クラスター数1での内部距離の0.7倍以下、外部距離に関しては、クラスター数1での内部距離の0.25倍以上とすることができる。それを満たすものの中で最小の数を最適値とすれば良い。もし、両

表-1 Thamas川流域での分類結果

algorithm	cluster number	elements of cluster center represented by data number
random K-means OF_1	1	10, 12, 15
	2	3, 6, 8, 13
	3	1, 2, 5, 7, 9, 14
	4	4, 11
random K-means OF_a	1	1, 3, 10
	2	2, 4, 5, 6, 7, 8, 9, 11, 14
	3	13
	4	12
DHB OF_s	1	9
	2	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14
	3	13
	4	15
ABF OF_n	1	1, 2, 3, 7, 8, 14
	2	4, 5, 9, 10, 11, 15
	3	6
	4	12, 13

表-2 サンプルデータの相関係数

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.00	0.86	0.59	0.71	0.90	0.22	0.93	0.42	0.51	0.92	0.37	0.94	0.89	0.78	0.45
2		1.00	0.34	0.56	0.68	-0.20	0.92	0.27	0.17	0.71	0.02	0.96	0.90	0.75	0.44
3			1.00	0.33	0.56	0.06	0.36	0.09	0.66	0.36	0.10	0.41	0.32	0.28	-0.22
4				1.00	0.49	0.52	0.81	0.09	0.79	0.83	0.79	0.62	0.45	0.42	0.30
5					1.00	0.38	0.72	0.73	0.35	0.84	0.27	0.77	0.88	0.91	0.64
6						1.00	0.14	0.46	0.58	0.51	0.84	-0.05	0.04	0.24	0.41
7							1.00	0.32	0.44	0.90	0.40	0.96	0.85	0.72	0.48
8								1.00	-0.16	0.55	0.16	0.37	0.66	0.78	0.87
9									1.00	0.56	0.76	0.27	0.08	0.08	-0.15
10										1.00	0.65	0.83	0.80	0.74	0.59
11											1.00	0.19	0.07	0.09	0.19
12												1.00	0.93	0.74	0.45
13													1.00	0.92	0.69
14														1.00	0.85
15															1.00

図-7 クラスター数と内部距離、外部距離、AICの関係

制約が同時に満足されないときは、内部距離に優先権があるとして、外部距離の制約ははずされる。内部距離も満たされないときは、適切な分類数が存在しなくなり、内部距離の制約を変えなければならない。但し、分類目的によっては、クラスター数に制限があり(小尻 他, 1989), 最適数が用いられない場合もある。

VI. 結 語

本研究で得られた成果を要約すると、次のようになる。

1) 目的関数の特性把握

OF_s と OF_n が系列間の Euclidean 距離の二乗和によって属性を決めるのに対して、 OF_a , OF_f は形状の評価を対象にしており、水文データの分類に適している。

2) 分類化手法の適合性

図-8 多目的最適化による TC と IC

最大化演算を含む目的関数には random K-means が適している。 DHB は正規化を含む二乗和型の目的関数と適合する。 ABF は非線形性の少ない Euclidean 距離の二乗和型に適している。

3) 水文データでのパターン分類化手順

水文データのパターン分類には、目的関数 OF_a と random K-means を使用すべきであろう。その結果、洪水時のピーク流量評価や降雨と流量、気温など、分類の目的にあわせて評価項目を変えることもできる。

4) クラスター数の決定

クラスター数の決定にあたっては、制約法、多目的最適化、AIC の導入を提案した。後 2 者で同じ結果が得られているので、それらをもとに制約法の条

件を決めるのが妥当であろう。AIC は計算量が少なく、単独でも適用が容易であった。

引用文献

- 1) Ismail, M.A. and Kamel, M.S. (1988): Multidimensional Data Clustering Utilizing Hybrid Search Strategies, 1988, Univ. of Waterloo, under Publishing
- 2) Kojiri, T., Ikebuchi, S. and Hori, T. (1986): Optimal Planning of Flood Control System Based on Risk Analysis of Flood Inundation, 4th Symp. of Flood Frequency and Risk Analysis, Louisiana, USA, pp.1-15
- 3) 小尻利治・藤井忠直(1989): 代表流況を用いた貯水池の実時間操作, 1989年土木学会中部支部年次講演会概要集, pp.182-183
- 4) Tou, J.T. and Gonzalez, R.C. (1974): Pattern Recognition Principle, Addison-Wesley Publishing Company, Applied Mathematics and Computation No. 7, pp.94-109
- 5) 宇治橋康行・高橋信忠・上谷昌史(1988): パターン認識概要による旬降水量時系列の解析とシミュレーション, 土木学会第33回年次講演会 II, pp.50-51
- 6) Unny, T.E., Panu, U.S. and Macniees, C.D (1981): Pattern Analysis and Synthesis of Time-dependent Hydrologic Data, Advances in Hydrosience, Vol. 12, pp.195-291

(1989・2・20受付, 1989・3・27受理)

小尻利治

昭和49年 京都大学大学院工学研究科修士課程終了
 昭和49年 京都大学工学部助手 (土木工学科)
 昭和56年 京都大学防災研究所助教授 (水資源研究センター)
 昭和61年 岐阜大学工学部助教授 (土木工学科)
 京都大学工学博士

Tharakkal E. Unny

1958 Univ. of IIT, 修士課程終了
 1960 Univ. of Dresden, 博士課程終了
 1963 Research Associate, Univ. of Shuttgart
 1968 Associate Professor, Univ. of Waterloo
 1971 Full Professor, University of Waterloo
 Dr. Engineering