

水文時系列の最適分類に関する研究

金沢大学工学部 学生員 松田 千 周
 同 上 正 員 高 瀬 信 忠
 同 上 正 員 宇 治 橋 康 行

1. まえがき

水文時系列データの模擬発生へのパターン認識の概念は、Panuらによって始めて導入され、その後、わが国の水文データの特性に合わせた拡張法が著者らにより提案された。また、小尻らは流量時系列のパターン分類に基づく過水状況の評価法を提案している。このようにパターン認識の概念は、水文学の分野において広く応用されつつある。しかしながら、最適パターン分類の基準あるいは最適分類を行うために用いるべきアルゴリズムおよびそのパラメータについては必ずしも明らかではない。本研究では、わが国の月降水量データを対象とし、ISODATAアルゴリズムを用いた水文時系列データの最適分類について検討する。

2. ISODATAアルゴリズム

パターンの認識・分類には大きく分けて2つの場合がある。一つは、与えられたパターンを特定のパターンクラスに割り当てる場合であり、文字認識あるいは音声認識等に見られる。他の一つは、パターンクラスを構成し同時に分類も行なう場合であり、クラスタリングと言われるものである。水文データのパターン分類は、後者のクラスタリングである。ここで用いるISODATAアルゴリズムは、BallとHallによって提案された、ユークリッド距離を類似度とする発見的なパターン分類アルゴリズムであり、パターン空間の構造が未知である場合に有効な分類法である。しかしながら、ISODATAのアルゴリズムは複雑であり、表1に示すように与える必要のあるパラメータ数も多く、データの持つ簡単な統計量からパラメータを定める基準は無いのが現状である。したがって、目的に応じた最適な分類結果を得るための、最適分類基準および最適パラメータの選択基準が必要とされている。

3. 最適分類の評価基準

パターン分類結果は、distance table あるいは variance table 等を用いて評価されるが、高次のパターンベクトルを視覚化することが困難であるため、パターン分類結果の評価は必ずしも容易ではない。ここでは、著者らが提案したパターン解析結果に基づくデータの模擬発生モデルへの適用を前提とし、以下に述べる2つの条件を満たす分類を最適分類と定義する。

1) ISODATAアルゴリズムによる分類結果の中で、分類された各パターンクラスの広がりを表す目的関数

$$J = \sum_{i=1}^{N_j} \sum_{x \in s_j} \|x - m_j\|, \quad m_j = \frac{1}{N_j} \sum_{x \in s_j} x \quad (1)$$

の値が小さくなる分類結果を3、4ケース選ぶ。ここに、 x はパターンベクトル、 N_j は第 j パターンクラスに属するパターンベクトルの数、 $\| \cdot \|$ はベクトルのノルムである。

2) 1) で選ばれた分類結果の中で、各パターンクラスのパターン内構造が多変数正規分布となり、データの模擬発生結果において、時系列レベルおよび月レベルにおいて2次までのモーメントが再現できる分類を最適分類と定義する。

4. 実データでの検討

ISODATAアルゴリズムの最適パラメータの検討に用いたデータは、金沢(100年間)、福井(8

7年間)、富山(46年間)、および福岡(94年間)の月降水量データである。表1に示された8個のパラメータの内、最適化を行ったパラメータは、これまでの研究から分類結果に大きく影響するパラメータであることが明らかにされている、 M 、 M_s 、 σ 、および δ の4個である。残る4個のパラメータは、分類結果に与える影響が小さいので、表2に示した値を固定して用いた。解析に用いた4地点のデータに対して定められた最適パラメータを表2に示す。表3には、パターンベクトルの各要素の分散の最小値 $\sigma_{\min}$ および各シーズンの平均ベクトル間の距離の最小値 $\delta_{\min}$ を示す。

表-1 ISODATAに用いるパラメーター

N_c	初期クラスター数
Z_0	初期パターンクラス中心
M	望むクラスター数
M_s	パターンクラス内の最小標本数
σ	標準偏差パラメーター
δ	最小クラスターペア間距離パラメーター
L	集中化されるクラスターペアの数
$I_{\max}$	最大反復計算回数

表-2 最適パラメーター

	パラメーター							
	N_c	Z_0	M	M_s	σ	δ	L	$I_{\max}$
金沢	1	(0.0,0.0,0.0)	9	10	40.0	70.0	1	200
富山	1	(0.0,0.0,0.0)	9	6	36.0	83.0	1	200
福井	1	(0.0,0.0,0.0)	9	15	40.0	70.0	1	200
福岡	1	(0.0,0.0,0.0,0.0)	9	5	33.0	83.0	1	200

表2から、望むクラスター数は各地点とも等しく9個であることが分かる。必要パターン数は、5~16個と幅があり、シーズン分割を通じて得られたパターンベクトルの総数との関連も明確ではなく一定値に固定することはできない。しかしながら、 M_s は始めに15程度の値を与え、前述の条件1)および2)を満たす分類が得られるまで、順次減らして行くことにより最適値を得ることが出来る。標準偏差パラメータ σ およびクラスター間の最小距離パラメータ δ は、表2および表3から、 $\sigma \approx \sigma_{\min}$ 、 $\delta \approx \delta_{\min}$ であることが分かる。したがって、 σ および δ の値は、シーズン分割に基づいてパターンベクトルを構成すれば、パターンベクトルの持つ統計量から最適値を定めることが出来る。

5. 結 語

わが国の4地点の月降水量データを用いた検討により、ISODATAアルゴリズムによる水文データの最適パターン分類の基準と最適パラメータ選択の基準を与えることが出来た。最適パラメータおよび最適値ないパラメータを用いた場合の模擬発生結果等の詳細については講演時に述べる。

表-3 パターンベクトルの統計量

	$\sigma_{\min}$	$\delta_{\min}$
金沢	37.4	71.5
富山	35.9	83.7
福井	44.2	70.4
福岡	33.3	85.4