数理解析研究所講究録
第229巻 1975年 99-109

柔かいオートマトンの実験的研究

東芝総研　森　健一

§1. はじめに

パターン認識の問題には，繁々，入力パターンが有限個の記号系列で表現され，これをいくつかのカテゴリーに識別する問題が出現する。パターンは本来多次元的な情報形態であり記号系列によって必ずしも完全に表現されるものではないが，適当な変換によってパターン認識の問題を記号系列の識別問題におきなおすことができる場合がある。

記号系列を識別するのに有限オートマトンを用いることが考えられるが，パターン認識の問題では入力パターンに雑音が必ず附随している入力記号系列が乱されることが多く，あるカテゴリーに属する全ての記号系列があらかじめ既知のときのみ有効な有限オートマトンの設計法ではほとんど役に立たない。また入力パターンに附随する雑音は多くの場合入力パターンに依存関係にあるため確率オートマトンで表現す

ることも難かしい。パターン認識における記号系列を識別するオートマトンは，Context Sensitive であり，雑音による入力記号系列の乱れに対して柔軟性をもたねばならない。また実用的な意味からは識別オートマトンの設計が容易であり，識別オートマトンの動作が速く，コンパクトに構成されねばならない。

筆者は昭和39年に自由手書郵便番号自動読取区分機を開発する際に上記のような考察から，オートマトンに「柔軟さ」を導入することを試みた。(1) 従来の有限オートマトンでは初期状態から出発して入力記号系列によって内部状態を遷移したとき，特定の最終状態に到達したか否かによって，そのオートマトンが入力記号系列を受容するか否かを決定するが，雑音によって入力記号系列が乱されると最終状態に到達できないことが多くなる。この欠点(「固さ」)を柔らげるため，内部状態の遷移に「ペナルティ」を考え，予期しない記号が入力記号系列の途中に出現したときにはその遷移にペナルティを課し，各カテゴリー毎に作られた識別オートマトンの中で入力記号系列に対し最もペナルティの少なかったオートマトンがその入力記号系列を受容することにした。この場合，入力記号系列を受容したオートマトンは必ずしも特定の最終状態に到達していなくともよい。

各カテゴリーの識別オートマトンには閾値が設けられており もし入力記号系列が遷移の途中でこの閾値を越えるならばその時点で特定の識別オートマトンの受容が拒否されるので，各カテゴリーの識別オートマトンのいずれに受容されるかを検査する手数が著しく少なくなり，速度が速くなる。

ペナルティを導入した識別オートマトンを特徴抽出過程，識別過程に用いたところ非常に有効であることが実証された。[2] 最近になってこのペナルティを導入したオートマトンをモデル化したペナルティ・オートマトンに関する論文[3]が発表されたのを機会に，ペナルティ・オートマトンを設計する際に必要な記号系列間の類似度とこれを用いた記号系列のクラスタリングおよびペナルティオートマトンの設計法について考察したい。

§2. 記号系列間の類似度

記号系列間の類似度については Misspelling を解析する計量言語学の分野で多くの研究がなされている。[4] まず予備的な定義として，記号系列とは非空の記号集合 Σ の要素を有限個並べたものを意味し，2つの記号系列 σ, ϕ があるとき

$$\sigma = \sigma_1 \sigma_2 \cdots \sigma_i \cdots \sigma_m , \quad \sigma_i \in \Sigma$$
$$\phi = \phi_1 \phi_2 \cdots \phi_j \cdots \phi_n \quad \phi_j \in \Sigma \qquad (1)$$

記号系列 $\psi = \psi_1 \psi_2 \cdots \psi_p$ が記号系列 σ および ϕ の共通部分

記号系列であるためには，少なくとも1組の添字の系列，

$i_1, i_2, \cdots i_p$ および $j_1\ j_2 \cdots j_p$ があって

(a) $p \leq \min(m, n)$

(b) $1 \leq k < \ell \leq p$ に対し，$i_k < i_\ell$ および $j_k < j_\ell$

(c) $1 \leq k \leq p$ に対し $\psi_k = \sigma_{i_k} = \phi_{j_k}$

が成立たねばならない。さらに $m \times n$ 次元のマトリックス C を考え，その要素 c_{ij} を

$$c_{ij} = \begin{cases} 1 & (\sigma_i = \phi_j) \\ 0 & (\sigma_i \neq \phi_j) \end{cases} \quad (1 \leq i \leq m,\ 1 \leq j \leq n) \quad (2)$$

としたとき，この C を一致マトリックスを呼ぶことにする。一致マトリックスの要素 c_{ij} で $i-j = a$（定数）を満す要素の集合を a 次対角と呼び，要素 c_{ij} から a 次対角への距離を $|a-(i-j)|$ で定義する。

記号系列間の類似度を定義するには，一致マトリックスの各要素における重み関数，一致マトリックスから最長の共通部分記号系列を選択する関数および共通部分記号系列から類似度を計算する関数をきめる必要がある。(4) パターン認識の場合，標準記号系列に全体としてどの程度類似しているかを評価することが必要であるため，重み関数として平均的な次数の対角に対する近かさによって評価することが適切である。(5)

選択された共通部分記号系列の平均対角次数は $\frac{1}{p}\sum_{\ell=1}^{p}(i_\ell - j_\ell)$

となるので重み関数を次式で定義する。

$$W_{i_k j_k} = C_{i_k j_k} \times \left\{ 1 - \frac{1}{\max(m,n)} \left| (i_k - j_k) - \frac{1}{P} \sum_{\ell=1}^{P} (i_\ell - j_\ell) \right| \right\} \quad (3)$$

但し. i_k, j_k は共通部分記号系列の k 番目の記号に対応するもとの2つの記号系列の記号の添字を意味する。

一致マトリックスから最長の共通部分記号系列を選択する関数は; 一致マトリックスのある1の値をもつ要素 C_{ij} に対応する記号を選んだとき, 次の記号として $k > i$, $\ell > j$ で $C_{k\ell} = 1$ となる要素に対応する記号を選んで構成した共通部分記号系列の中で最も長い系列を選ぶ。

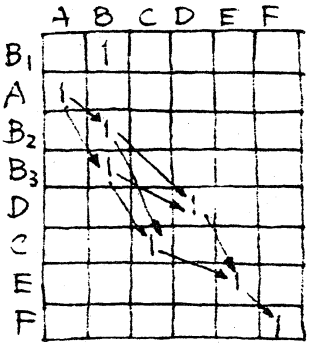
例1. $\Sigma = \{A, B, C, D, E, F\}$

$\Upsilon = ABCDEF$

$\Phi = BABBDCEF$

とすると最長の共通部分記号系列は

4通りある。それを E であらわすと



一致マトリックス

$$E = \{AB_2DEF, AB_3DEF, AB_2CEF, AB_3CEF\}.$$

ある最長の共通部分記号系列 Ψ に対し. その構成要素の一致マトリックス上での重み関数を (3) 式とすると, 2つの記号系列 Υ および Φ に関して共通部分記号系列 Ψ を介しての類似度を次式で定義する。

$$S(\Upsilon, \Phi / \Psi) = \left\{ 1 - \left| \frac{\frac{1}{2}(m-n) - \frac{1}{P} \sum_{\ell=1}^{P} (i_\ell - j_\ell)}{\max(m,n)} \right| \right\} \cdot \sum_{k=1}^{P} \frac{W_{i_k j_k}}{\max(m,n)} \quad (4)$$

さらに2つの記号系列 σ および ϕ の類似度を次式で定義する。

$$S(\sigma,\phi)=\max_{\psi\in E}\{S(\sigma,\phi/\psi)\} \qquad (5)$$

2つの記号系列間の類似度 $S(\sigma,\phi)$ は

$$0\leq S(\sigma,\phi)\leq 1 \qquad (6)$$

となり，記号系列 σ と ϕ の共通部分記号系列が空のとき 0 となり，記号系列 σ と ϕ が一致するとき 1 になる。

例2. 例1の問題に対し，類似度を計算する。

$\psi_1=AB_2DEF$, $\psi_2=AB_3DEF$, $\psi_3=AB_2CEF$, $\psi_4=AB_3CEF$

$S(\sigma,\phi/\psi_1)=0.558$

$S(\sigma,\phi/\psi_2)=0.543$

$S(\sigma,\phi/\psi_3)=0.517$

$S(\sigma,\phi/\psi_4)=0.519$

$\therefore\ S(\sigma,\phi)=0.558$

上の定義では記号間の類似度は1か0で評価した。すなわち，$A,B\in\Sigma$ とすると $S(A,A)=1$. $S(A,B)=0$. しかしながら，パターン認識の問題では記号間の類似度が全ていついて1か0になる場合は少なく，異なる記号間の類似度が適当な値をもつ場合が多い。この場合に記号系列間の類似度の定義を拡張することは容易で，まず一致マトリックス

の要素 C_{ij} を記号系列 σ の i 番目の記号 σ_i と，記号系列 ϕ の j 番目の記号 ϕ_j の類似度で置直し，(2) に代って

$$C_{ij} = S(\sigma_i, \phi_j) \quad , (1 \leq i \leq m, 1 \leq j \leq n) \qquad (2)'$$

とする。共通部分記号系列の選択は0でない要素の組合せの全てについて行なう必要がある。重み関数と類似度の評価式は (3), (4), (5) 式を用いる。

例3. 例1に $S(A,B)=0.6$, $S(C,D)=0.7$ を加えた場合の記号系列 ABCDEF と BABBDCEF の類似度を計算する。

$S(\sigma,\phi/AB_3DCEF) = 0.583$

$S(\sigma,\phi/AB_2DEF) = 0.558$

$S(\sigma,\phi/B_2B_3DCEF) = 0.523$

$S(\sigma,\phi) = 0.583$

ϕ \ σ	A	B	C	D	E	F
B_1	0.6	1				
A	1	0.6				
B_2	0.6	1				
B_3	0.6	1				
D			0.7	1		
C			1	0.7		
E					1	
F						1

§3. 記号系列のクラスタリング

識別オートマトンを設計する際にあるカテゴリーに属する記号系列のサンプルが与えられたとき，サンプル全体を代表する記号系列はどれかをきめたいときがある。すなわちあるカテゴリーに属する標準記号系列をサンプル記号系列群からきめようとする問題がある。さらにサンプル全体を1つの記

号系列で代表させることが困難なときは，サンプル記号系列群を適当な数の類似群にクラスタリングする問題を考えることができる。

今サンプル記号系列の集合をGとし，Gに属する記号系列の数をNとするとき

$$S_\sigma = \frac{1}{N-1} \sum_{\substack{\phi \in G \\ \phi \neq \sigma}} S(\sigma, \phi), \quad \sigma \in G \qquad (7)$$

を計算し，S_σが最大となる記号系列をサンプル記号系列Gの代表記号系列という。

サンプル記号系列Gを任意の数M個の類似群にクラスタリングする手順は，多次元ベクトルで表現されたサンプルデータ群をユークリッド距離を評価関数として漸近的にクラスタリングする手法[6]に準じて，次の通りとすることができる。

記号系列群のクラスタリングの手順

(1) サンプル記号系列の集合Gから適当なM個の記号系列を核記号系列としてえらぶ，

(2) Gに属する任意の記号系列をとり出し，核記号系列との類似度を計算し，最も大きな類似度を示す核記号系列の群にその記号系列は属するとして分類する。Gの全ての記号系列を分類し，M個の群に分ける。

(3) 各分類群の代表記号系列を計算し，これを新しい核記

号系列とする。

(4) 各分類群の代表記号系列が変化しなくなるまで(2)(3)の手順を繰返す。もしある分類群で複数個の記号系列が同じ最大の S_σ をもつときは、他の分類群の代表記号系列に平均的に最も類似していない記号系列を代表記号系列とする。

例4 $G=\{BA, BB, BBA, ABB, ABA, AABA\}$ の代表記号系列を求め、次に2つの類似群に分類する。

まず記号系列間の類似度を計算すると右表のようになり、これから代表記号系列は ABA である。最も類似していない BB と AABA を最初の核記号系列とすると (BB), (ABA, BA, BBA, ABB, AABA) の2群に分類される。最も類似した ABB と ABA を最初にとっても同じ結果が得られる。しかし、BBA と ABA を最初にすると (BBA, BA, BB), (ABA, BA, ABB, AABA) に分割される。

	BA	BB	BBA	ABB	ABA	AABA
BA	1.0	0.5	0.56	0.28	0.56	0.38
BB	0.5	1.0	0.56	0.56	0.28	0.25
BBA	0.56	0.56	1.0	0.44	0.67	0.44
ABB	0.28	0.56	0.44	1.0	0.67	0.44
ABA	0.56	0.28	0.67	0.67	1.0	0.00
AABA	0.38	0.25	0.44	0.44	0.00	1.0

最初にえらぶ核記号系列によってふつうは常に同一のクラスタリングがなされないが、分類された結果はほぼ似た類似群が構成されるという意味で安定したクラスタリングである。

§4. ペナルティ・オートマトンの設計法

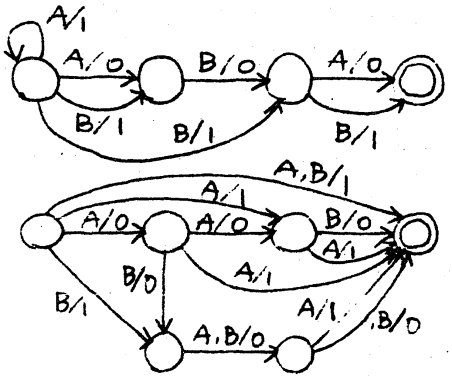
文献(3)のペナルティオートマトンの構成法をもとにし，記号系列間の類似度を用いて並列ペナルティオートマトンを設計する手順は，次のようになる。

(1) 各識別カテゴリー毎にサンプル記号系列から代表記号系列を求め，ペナルティ0のペナルティオートマトンを構成する。

(2) 代表記号系列との間の類似度が大きい順にサンプル記号系列を受容するようにペナルティ1の遷移を付加してゆく。同時に他のカテゴリーの記号系列を受容しないようにチェックする。

(3) あるカテゴリーに属する記号系列が自分の代表記号系列より他のカテゴリーの代表記号系列により類似しているときは，そのカテゴリーを2つにクラスタリングし新しい遷移ルートをペナルティ1で付加する。

例5 (BA, BB, BBA, ABB, ABA, ~~AABA~~), (A, B, AA, AB, AAA, AAB, BAA, BAB, BBB, ABBA) をサンプル記号系列とする並列ペナルティオートマトンを構成する。



§5 おわりに

記号系列の識別オートマトンを構成するのに必要な記号系列間の類似度および記号系列のリラスタリングの手順について述べ，並列ペナルティ・オートマトンの設計への応用を試みた。並列ペナルティ・オートマトンの理論や構成法についてはまだ多くの研究すべき事項が残っている。

参考文献

(1) H.Genchi, K.Mori, S.Watanabe and S.Katzuragi, Proc IEEE, 56,8, pp 1291~1301, (1968).

(2) k.Mori, H.Genchi, S.Watanabe, S.Katsuragi, Pattern Recognition. 2, pp 175-185, (1970).

(3) 阿部圭一, 電通学会研究会資料. PRL74-5, pp 45-54, (1974).

(4) C.N.Alberga, J.ACM, 10, 5, pp 302-313, (1967)

(5) R.Vivès and J.Y. Gresser, Proc. IJCPR, pp 308-317, (1973).

(6) G.H.Ball, Proc. FJCC, Vol 27, (Spartan Books), pp 533-559, (1965).