数理解析研究所講究録
229巻 1975年 145-155

情報圧縮と文法推論

九大 理 情報研 有川節夫

文法推論による情報圧縮とその可能性について考え，さらに，右延長文法とその推論について考える。

1. 情報の圧縮と損失

一般に情報処理の過程はある目標 θ に沿った何段階かのデータの変換からなるといえる。データ $\mathcal{D}$ を記録するのに要する面積を $|\mathcal{D}|$ とするとき，$|\mathcal{D}| > |T(\theta, \mathcal{D})|$ であれば，$\mathcal{D}$ は変換 T によって圧縮されたことになる。変換には，検索効率を高めるためにデータを組織化する場合，あるいは，データを図式化する場合のように $|\mathcal{D}| \leq |T(\theta, \mathcal{D})|$ となる変換もあるが，多くの変換は圧縮を指向したものであるといえよう。

また，変換には，とくに，圧縮の場合には，"情報の損失"を伴うことが多い。ここで情報の損失の定義が問題になるがそれは，情報処理過程に対する目標 θ を介してつぎのように把えられる：即ち，変換 T が目標 θ の下でデータ $\mathcal{D}$ に対して情報の損失を伴わないとは，$(\theta, \mathcal{D})$ に適用可能な任意の変

損 R に対して

$$R(\theta, D) = S(T(\theta, D)) \tag{1.1}$$

となる変換 S が存在することである。こうすると，一見大幅な圧縮をする変換でもその目標のためには，情報の損失は無いことがあることなども説明できる。

一般には，情報の損失を伴わない大幅な圧縮は目標が単純明確である場合には可能であることが多いが，多目的に使用されるファイルの作成，あるいは用途の細目が完全には分っていないデータの整理などの場合には困難である。

2. 文法推論による圧縮

このような局面に対処する1つの方法として，θ に依存しない情報の圧縮，損失を考える，即ち，"任意の θ の下で情報の損失を伴わない圧縮" を考える方法がある．この場合には，変換は単に $T(D)$ の形で扱われ，情報の損失に関する式 (1.1) は

$$D = S(T(D)) \tag{2.1}$$

で置き換えられる．"T が D に対して情報の損失を伴わない" とは "$T(D)$ から D を復元することが可能である" ということになる．このように θ を無視することにより，圧縮率の減少はまぬかれないが，一般的な議論の可能性が生れる。

一般的な1方法として，ここでは，文法推論の方法を利用

することを考える。文法推論とは，与えられたサンプル集合から，その特徴を抽出して，それを文法あるいはオートマトンとして記述する方法である．

ところで，データは記号列の系列または集合として把えられるが，ここでは単に記号列の有限集合であると考える。即ち，Σ を記号の有限集合とするとき，データ D は Σ^* の有限部分集合であるとする．

文法推論による圧縮法とは，データ D に対して，文法族 $\mathfrak{G}=\{G_1,\cdots,G_n\}$ と述語 P とを条件

$$D = L(\mathfrak{G}, P) \qquad (2.2)$$

$$|D| > |\mathfrak{G}| + |P| \qquad (2.3)$$

を満たすように決定する方法である。ここで，文法族，述語の選択範囲，L の定義が問題になる．文法に関しては，効率よい推論が行えるために，有限オートマトンかそれに類したもの，述語はスマリヤンの rudimentary attribute 的なものが妥当であろう。L はその際に自然に定まる．

しかし，当面は現在の文法推論法を利用するために，

$$D \subseteq L(G(D)) \qquad (2.4)$$

なる文法 $G(D)$ を推論して，適当なフィルターとしての述語 P を選び，余分な $L(G(D)) - D$ を振り落して

$$D = L(G(D), P) \equiv \{x \in L(G(D))\ ;\ P(x)\} \qquad (2.5)$$

とすることを考えよう。たとえば，データ

$$\mathcal{D}=\{a,\ aab,\ aabab,\ \cdots,\ a(ab)^{100}\} \tag{2.6}$$

に対して，文法と述語

$$G(\mathcal{D})=\{S\rightarrow aA,\ A\rightarrow abA,\ A\rightarrow \varepsilon\} \tag{2.7}$$

$$P(x)\leftrightarrow |x|<202 \tag{2.8}$$

を仮定すれば，$\mathcal{D}=L(G(\mathcal{D}),P)$ となる。この場合の変換 T は，$\mathcal{D}$ から $G(\mathcal{D})$ と P を作る操作である。データの面積は区切り記号を無視すれば，$|\mathcal{D}|=10201>19=|T(\mathcal{D})|$ となり，$\mathcal{D}$ は T によって情報の損失を伴うことなく $T(\mathcal{D})$ に圧縮されたことになる。

フィルターとしての述語 P は正確に $\mathcal{D}=L(G(\mathcal{D}))$ となるように $\mathcal{D}(G)$ を推論すれば不要であるが，$\mathcal{D}$ は有限であるから，$|G(\mathcal{D})|$ の値はあまり小さくならないことが多い。(上の $\mathcal{D}$ を有限オートマトンので表現すると状態は最少の場合でも 103 個必要でありそれを2進数でコード化すると約 600 のスペースが必要であり，推移表の面積は $600\times 2=1200$ となる)。このことと，$\mathcal{D}$ がある母集団からのサンプルである場合には，推論問題は付帯条件を除いた (2.4) だけで扱う方が自然であるので，先ず(2.4)を考え，更に述語 P を考えるのである。

3. 圧縮の可能性

圧縮の可能性，度合の問題はデータの複雑さの問題として

扱うことができる．データ D の複雑さ，すなわち，可能な圧縮の度合いは

$$H(D) = \min\{|G|+|P|;\ L(G,P)=D\} \tag{3.1}$$

で定義できる．定義から直ちに

$$H(D) \leq |D| \tag{3.2}$$

であることが分る。

多くの D に対して $H(D)$ の値が十分小さくなることを期待したいのであるが，悲観的材料として Kolmogorov の複雑さに関する結果がある．即ち，$x \in \Sigma^*$ の複雑さを $K(x)$ とすると，

$$K(x) < |x| \tag{3.3}$$

$$\#\{x;\ K(x) < l-m\} / \#\{x;\ |x|=l\} \leq 2^{-m+1} \tag{3.4}$$

であることが知られている．つまり，大多数の x に対して，情報の損失のない大幅な圧縮は期待できないことになる．しかも，D を適当に順序付けられた系列であると考えれば，一般に

$$K(D) \leq H(D) \tag{3.5}$$

であるから，任意のデータを文法推論による圧縮法により，大幅に圧縮することは不可能である。

しかし，集合 D の要素の順序付けの仕方によって $H(D)$ の値が小さくなる可能性はある．また，D に明確な規則性がある場合，あるいは D についての付加的情報が予め分っている場合には大幅な圧縮も可能である。一方，再び Kolmogorov の理論から，$K(x)$ の値が大

きければ大きい程 x は random に近いことが知られているので、$H(D)$ の値を下げることが不可能である場合には、D は random と考えて差しつかえなく、その分布も確定するので、D について考えられる統計的問題はすべて解決されることになる。この場合は、その分布を記録するだけで十分であるので、再び大幅な圧縮への道が開ける。勿論ここの意味での情報の損失は伴うことになるが、D に関しては、D そのもの以外に、いかなる決定論的で有効な特徴抽出の手法もないのである。

4. 右延長文法とその推論

文法推論において、データの特徴を記述するシステムは、従来のオートマトン、文法に限定されたものではない。ここでは、推論、特徴抽出の最も素朴な形式の定式化とみられる文法種を導入し、1つの推論を考えることにする。

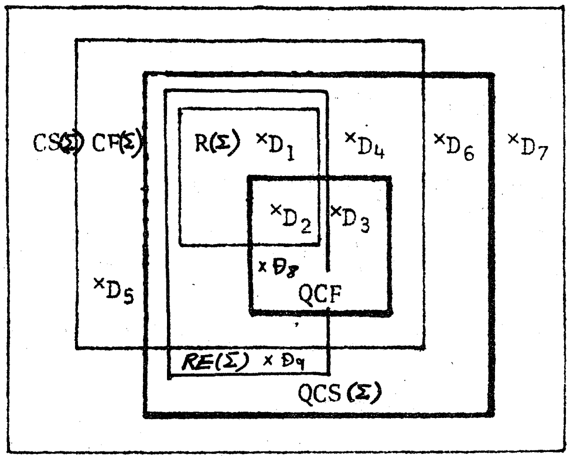
4.1. 擬文法とその性質

文脈依存文法 $G=(V, V_T, P, S)$ による sentential forms の全体のなす集合 $S(G)=\{w ; S \overset{+}{\Rightarrow} w\}$ を G によって生成される擬文脈依存言語 (qcsl) という。$\overset{+}{\Rightarrow}$ は1回以上のルール適用による導出を示す。qcsl について考える際には、V_T を明示する必要はないから、$G=(V, P, S)$ と書き、これを qcsg と呼ぶ。$G=(V, V_T, P, S)$ が文脈自由であるときは、qcfg という。

gcsg $G=(V, P, S)$ において，P の要素が $S\to\alpha$，$\alpha\to\alpha\beta$ の型であるとき，G を右延長文法といい，以後 $\underline{S}:=\{\alpha;\ S\to\alpha\in P\}$，$\underline{P}:=\{(\alpha,\beta);\ \alpha\to\alpha\beta\in P\}$ とする。

記号の集合 Σ 上の csl の全体を $CS(\Sigma)$，cfl の全体を $CF(\Sigma)$，正則集合の全体を $R(\Sigma)$，gcsl の全体を $QCS(\Sigma)$，gcfl の全体を $QCF(\Sigma)$，右延長言語の全体を $RE(\Sigma)$ とする。

$\#\Sigma\geq 2$ ならば，これらの言語族には下図のような関係がある：



$D_1=\{A\}\cup\{A^{2n+2};\ n\geq 0\}$，　$D_2=\{A\}^*$

$D_3=\{A^nBA^n;\ n\geq 0\}$，　$D_4=\{A^nB^n;\ n\geq 0\}$

$D_5=\{AB^nAB^{n+1};\ n\geq 1\}^*$，　$D_6=c(\underline{S}(G));\ L(G)=\{a^{2^n};\ n\geq 0\}$，

$D_7=\{A^{2^n};\ n\geq 0\}$，　D_8：$\{A,B\}$ 上の Dyck 言語，

ここに c は G のアルファベットを $\{a, S, A_1, \cdots, A_p\}$ とするとき，

$$c(a)=A^{p+1}B,\quad c(A_i)=A^{p-i+1}B^{i+1}\ (1\leq i\leq p),\quad c(S)=S$$

で定義される homomorphism である。D_9 については[14]を参照されたい。

なお，$\#\Sigma=1$ の場合には，これらの言語族間には，

$$QCF(\Sigma) \subsetneq R(\Sigma) = RE(\Sigma) = QCS(\Sigma) \subsetneq CS(\Sigma) \qquad (4.1)$$

の関係がある．$QCF(\Sigma)$ 等の他の性質について少し述べる．

<u>定理1.</u> $\#\Sigma \geq 2$ とする．$QCS(\Sigma)$, $QCF(\Sigma)$, $RE(\Sigma)$ はつぎのぎの演算の下でも閉じていない：(1) union, (2) product, (3) $*$, (4) intersection, (5) Σ^* に関する complementation, (6) homomorphism.

また，$QCS(\Sigma)$, $QCF(\Sigma)$ は正則集合との intersection の下でも閉じていない．

<u>定理2.</u> $\Sigma=\{A\}$ とする．

(1) $D \subseteq \Sigma^*$ が qcfl $\Leftrightarrow$ $D=\delta(D) \cup X\cdot\{A^r\}^*$ となる有限集合 $X \subseteq \Sigma^*$ と整数 r が存在する．

(2) $D \subseteq \Sigma^*$ が qcsl $\Leftrightarrow$ $D = F \cup X\cdot\{A^r\}^*$ となる有限集合 $X, F \subset \Sigma^*$ と整数 r が存在する．

(3) $QCF(\Sigma)$ は union の下で必ずしも閉じていないが，$D_1, D_2 \in QCF(\Sigma)$ が共に無限であれば，$D_1 \cup D_2 \in QCF(\Sigma)$ である．

<u>定理3.</u> $QCS(\Sigma, k)=\{D \subset \Sigma^*; (\exists \text{qcsg } G)[S(G)=D \ \& \ o(G)=k]\}$,

$$o(G) = \max\{|\alpha|; (\exists\beta)[\alpha\to\beta \in P]\}$$

とすると，(1) $QCF(\Sigma)=QCS(\Sigma,1)$, (2) $QCS(\Sigma,k) \subsetneq QCS(\Sigma,k+1)$, (3) $QCS(\Sigma)=\bigcup_{k=1}^{\infty} QCS(\Sigma,k)$ であり，さらに

(4) $D \in QCS(\{A\},k) \Leftrightarrow D=F\cup X\cdot\{A^r\}^*$ となる有限集合 $X, F \subset \Sigma^*$ と整数 r が存在し，各 $z\in F$ に対して $|z|<k$ となる．

4.2. 右延長文法の推論

前節でみたように擬文法による言語族は通常の言語族とは極めて異った性質をもつが，文法は文の生成のために何も補助的な記号，状態等を必要としないので推論には都合がよい．以下に右延長文法の推論について簡単にのべる．

長さの短かい順に並んだ n 個のサンプル系列の集合

$$D = \{x_1, x_2, \cdots, x_n\};\ |x_i| \leq |x_{i+1}|\ (1 \leq i \leq n) \tag{4.2}$$

が与えられた時点で条件

$$D \subseteq S(G) \quad \& \quad S(G)|_D = D \tag{4.3}$$

をみたすような右延長文法 G を推論する1方法を考える．ここに

$$M = max\{i;\ i \geq 1\ \&\ |x_i| < |x_n|\} \tag{4.4}$$

とするとき

$$S(G)|_D = \{x \in S(G);\ |x| \leq |x_M|\} \cup (\{x_i;\ M < i \leq n\} \cap S(G)) \tag{4.5}$$

である．推論の手順の詳細は省略するが，本質的な部分は，各 x_i に対して x_i を導出するような $x_j\ (j < i)$ を見つけることにある．x_j が見つかれば，$x_j \Rightarrow x_i$ を導く規則を検討して，D の長さまでの語で D に属さない語を導出しなければ，その規則を P の元として採用する．$x_j \Rightarrow x_i$ を導くすべての規則が D 以外の $|x_n|$ 以下の語を導出するか，x_j が見つからないときには，x_i を S の元として採用する．このようにして，右延長文法 $G = (\Sigma, P, S)$ を決定する．最後にこの方法による推論過程の状況を例示しておこう：

i	x_i	C	S	P
1	011		011	
2	00111	(0,01)	00111	
3	01101	(011,01),(01,10)	01101	
4	0010111	(00,10),(0,0101),(0,01),(001,01)		(00,10)
5	0011101	(00111,01),(0,01)		(111,01)
6	0110101	(01101,01),(0110,10),(011,0101),(01,1010),(01,10),(011,01)		(10,10)
7	001010111			
8	001011101			
9	001110101			
10	011010101			
11	00101010111			

5. おわりに

文法推論による情報圧縮について基礎的考察を行い，その可能性について論じた。また，右延長文法とその周辺の文法を考え，推論についての簡単な考察を試みた。これらの中には筆者のこの方面での研究の方針を述べただけのところも少くない。したがって，解決すべき多くの問題が残されている。データの定義，目標と呼んできたものの定式化，変換の定式化，付加的情報の扱い，推論手順の比較，改良などの問題がある。これらについては別の機会に論じるつもりである。

参考文献

1. Moore, E.F. Gedankenexperiments on sequential machines, in Automata Studies, Princeton Univ. Press (1956).

2. Solomonoff, R.J. A formal theory of inductive inference, Inform. Cont. 7 (1964), 1-22.

3. Gold, E. M. Limiting recursion, J. Symbolic logic 30 (1965), 28-48.

4. 榎本肇：サンプル記号列によるオートマトンの記述と構成，日科技連 数理計画シンポジウム「オートマトン」(1971) 141-163

5. Feldman, J. Some decidability results on grammatical inference and complexity, Inform. Cont. 20 (1972), 244-262.

6. Biermann, A. W. An interactive finite-state language learner, 1st US-Japan Comp. Conf. Proc. (1972), 13-20.

7. Banerji, R. B. Theory of problem solving, Amer. Elsevier (1969).

8. Smullyan, R. Theory of formal systems, Princeton Univ. Press (1961).

9. Tanatsugu, K. On inference of a concise description of finite and infinite concepts, RIFIS-RR 34 (1973), 1-13.

10. Zvonkin, A. K. & Levin, L. A. The complexity of finte objects and the development of the concepts of information and randomness by means of the theory of algorithms, Rus. Math. Surv. 25 (1970), 83-124.

11. Tanatsugu, K. & Arikawa, S. On characteristic sets and degrees of finite automata, RIFIS-RR 45 (1974), 1-14.

12. Arikawa, S. On the languages defined by sentential forms of context-free grammars, RIFIS-RR 17 (1970), 1-12.

13. Arikawa, S. Closure and nonclosure properties of quasi context-sensitive languages, RIFIS-RR 37 (1973), 1-25.

14. 大芝猛．有川：右延長文法による言語の族の大きさについて 京大数研講究録 213 (1974) 104-122.