制御集合にもとづく線型言語の帰納的推論

5P-2

高田裕志

富士通㈱国際情報社会科学研究所

1. 導入

言語はそれを生成する文法を特定することによって同定される。ここでは、線型言語L を同定する方法を論じる。はじめに、アルファベットΣ上の任意の線型言語Lが、Σ上のある固定された線型文法G^0とG^0の生成規則上の正則制御集合C から生成されることを示す。この事実から、G^0に対する正則制御集合C を同定することによって、線型言語L を同定する。言語の構造の情報とともに、L の所属に関するoracleとL を生成する線型文法G の補助的な情報が与えられたとき、L を正則集合の同定と同じ方法で同定できる。

2. 表現定理

Σを有限のアルファベット、Σ^* をΣ上のすべての記号列からなる集合とする。さらに、λを空記号列、$|u|$をu の長さとする。

$M=\langle K,\Sigma,\delta,q_0,F\rangle$を(決定性)有限状態受理機とする。ここで、K は状態の空でない有限集合、δは状態遷移関数、q_0は初期状態、F は最終状態の集合である。Σ^* の部分集合R を受理する有限状態受理機が存在するとき、R は正則であるという。R が正則ならば、R を受理する状態数が最小の有限状態受理機が同型を除いて唯一存在する。この受理機をR に対する正準(canonical) 有限状態受理機という。

$G=\langle N,\Sigma,\Pi,S\rangle$ を文脈自由文法とする。ここで、N は非終端記号の、Π は生成規則の空でない有限集合、S は初期記号である。 $N\cup\Sigma$をV で表す。G のどの生成規則も $S\to\lambda$、 $A\to a$、 $A\to aB$または $A\to Ba$ $(S,A,B\in N,\ a\in\Sigma)$ の形をしているとき、G を (標準形の) 線型文法という。特に、ただ1つの非終端記号S だけからなるとき、G は最小であるという。各生成規則にはラベルがつけられているものとし、それをπ_i で示す。また、G は無用な非終端記号を含まないと仮定する。

【定義】 Σを$\{a_1,a_2,\ldots,a_n\}$とする。このとき、生成規則の集合Π^0 が

$$\{\ S^0\to\lambda,\ S^0\to a_1S^0,\ S^0\to a_2S^0,\ \ldots,\ S^0\to a_nS^0,$$
$$S^0\to S^0a_1,\ S^0\to S^0a_2,\ \ldots,\ S^0\to S^0a_n,$$
$$S^0\to a_1,\quad S^0\to a_2,\quad \ldots,\ S^0\to a_n,\ \}$$

である最小線型文法$G^0=\langle\{S^0\},\Sigma,\Pi^0,S^0\rangle$ を万能(universal) であるという。

アルファベットΣに対して、万能最小線型文法G^0は唯一に存在する。

G を線型文法、$x_0\overset{\pi_1}{\Rightarrow}_G x_1\overset{\pi_2}{\Rightarrow}_G\cdots\overset{\pi_k}{\Rightarrow}_G x_k$ をG における導出(derivation)とする。ここで、x_{i-1} からx_i $(1\le i\le k)$ への遷移において生成規則π_i が適用されるものとする。$x,y\in V^*$, $\alpha\in\Pi^*$ に対して、$x\overset{\alpha}{\Rightarrow}_G y$ はG において導出 $x=x_0\overset{\pi_1}{\Rightarrow}_G x_1\overset{\pi_2}{\Rightarrow}_G\cdots\overset{\pi_n}{\Rightarrow}_G x_n=y$ が存在することを意味する。ここで、$\alpha=\pi_1\pi_2\cdots\pi_n$ である。αを導出$x\overset{\alpha}{\Rightarrow}_G y$の付随語(associate word)という。G によって生成された言語L(G)とは集合

$$L(G)=\{\ w\in\Sigma^*\mid S\overset{\alpha}{\Rightarrow}_G w,\quad \alpha\in\Pi^*\ \}$$

である。L=L(G)となる線型文法G が存在するならば言語L は線型であるという。

【定義】 線型文法 $G=\langle N,\Sigma,\Pi,S\rangle$ が次の2つの条件を満たすとき、正準であるという：

1. A≠B であるN の任意の要素A,B に対して、$A\overset{\alpha}{\Rightarrow}_G w$ であるが $B\overset{\beta}{\Rightarrow}_G w$ でない記号列 $w\in L(G)$が存在する。
2. $A\to aB$, $A\to aC$または $A\to Ba$, $A\to Ca$ $(B\neq C)$である生成規則のペアは存在しない。

特に、条件2を満たすときG は決定的であるという。

【定義】 C をΠ^* の任意の正則な部分集合とする。このとき、

$$L_C(G)=\{\ w\in\Sigma^*\mid S\overset{\alpha}{\Rightarrow}_G w,\ \alpha\in C\ \}$$

を正則制御集合C とG によって生成された言語という。

《定理1》 任意の線型言語L に対して、$L=L_C(G^0)$となる正則制御集合C が存在する。□

証明は省略する (文献[3]を参照、以下の定理、命題についても同様である)。方針としては、G^0に対する正則制御集合C を受理する有限状態受理機の各状態を、L を生成する線型文法G の非終端記号に対応させる。定理1の逆の場合も成り立つ。

《定理2》 G^0をアルファベットΣ上の万能最小線型文法、C をG^{0}に対する正則制御集合とする。そのとき、$L=L_C(G^0)$は線型言語である。□

Inductive Inference of Linear Languages based on Control Sets

Yuji TAKADA (yuji@iias.iias.fujitsu.junet)

International Institute for Advanced Study of Social Information Science(IIAS-SIS), FUJITSU LIMITED

定理1と2から、線型文法G が与えられたとき、L(G)=$L_c(G^\circ)$となるC を受理する有限状態受理機M を対応させることができる。

《命題1》 任意の線型文法G に対して、L(G)=L(G')となる決定的線型文法G'が存在する。□

《系》 任意の線型言語L に対して、L=L(G)となる正準線型文法G が存在する。□

任意の線型言語L に対してL を生成する正準線型文法G が存在する。しかし、その文法は唯一ではないので有限状態受理機における「正準」と線型文法における「正準」は必ずしも同義ではない。

《命題2》 G を線型文法、M をC=T(M)かつL(G)=$L_c(G^\circ)$であるG に対応する有限状態受理機とする。このとき、G が正準ならばM は正準である。□

一般に、逆の場合は成り立たない。

このように、線型言語L を同定するには、L を生成する正準線型文法G に対応する正準有限状態受理機M を同定すればよい。

3. 線型言語の同定アルゴリズム

同定アルゴリズムは言語の構造に関する情報が利用可能であるとする。言語の構造はかっこ文法(parenthesis grammar) によって記述することができる。

【定義】 線型文法G=<N, Σ, Π, S> のかっこ文法<G> はΠのすべての生成規則 A→x (x∈V*) を A→<x> で置き換えることによって得られる。ここで、"<",">" はΣの要素でない特別の記号である。

W を記号列の集合とする。このとき、Pr(W) はW の要素のすべてのprefixの集合である。Pr(W) は空でないとき、λを含む。X とY が集合であるとき、 X⊕Y でX とY の対称差を表す。

【定義】 G を正準線型文法、<G>=<N, Σ, Π, S> をG のかっこ文法、G°をΣ°上の万能最小線型文法、<G°>=<{S°}, Σ, Π°, S°> をG°のかっこ文法とする。L(<G>)の表現標本(representative sample) Raとは<G> のすべての生成規則πに対してπが導出$S \overset{\alpha}{\Rightarrow}_{<G>} w$ に現れるRaの要素w が存在するL(<G>)の有限部分集合である。Raに関するC の付随表現標本Rcとは集合

$$Rc = \{\alpha \mid S0 \overset{\alpha}{\Rightarrow}_{<G>} w,\ w \in Ra\}$$

である。

言語の構造を与えることによって導出を特定することができる。したがって、L(<G>)の要素w に対して、w の<G°>における導出を表す付随語αが一対一に対応する。

線型言語同定アルゴリズムLID は表現標本RaとL(<G>)の所属に関するoracleから有限状態受理機M を同定する。LID はAngluin [1]によるアルゴリズムIDを線型言語版に修正したものである。

アルゴリズムLID において、Pcの各要素は可能な状態を表す。d_0は「落とし穴」状態を表す。関数f は任意の$\pi^\circ \in \Pi^\circ$に対して$f(d_0, \pi^\circ)=d_0$、各$\mu \in Pc$に対して$f(\mu, \pi^\circ)=\mu\pi^\circ$ と定義される。f はM の遷移を表す。関数g は付随語αに対して$S^\circ \overset{\alpha}{\Rightarrow}_{<G>} s$ となるV*の要素s を対応させる。LID はPcの要素を同値類に分割する。その同値類の各要素がM の状態となる。

アルゴリズム LID
入力であるRaから次の集合を構成する
$Rc = \{\alpha \mid S^\circ \overset{\alpha}{\Rightarrow}_{<G>} w,\ w \in Ra\}$
$Pc = Pr(Rc)$
$Pc' = Pc \cup \{d_0\}$
$Tc' = Pc' \cup \{f(\mu, \pi^\circ) \mid \mu \in Pc,\ \pi^\circ \in \Pi^\circ\}$
$Tc = Tc' - \{d_0\}$
ステップ 0
すべての$\mu \in Tc'$ に対して$E_0(\mu)$ を空にする
iを1、ν_1 をλにする
ステップ i
$E_i(d_0)$を空にする
各$\mu \in Tc$に対して記号列$g(\mu\nu_i)$をL(<G>)のoracleに質問し、$E_i(\mu)$ を構成する
$g(\mu\nu_i) \in L(<G>)$ならば$E_i(\mu)$ を$E_{i-1}(\mu) \cup \{\nu_i\}$ とし、そうでなければ$E_{i-1}(\mu)$ とする
$E_i(\mu)=E_i(\mu')$だが$E_i(f(\mu, \pi^\circ))\neq E_i(f(\mu', \pi^\circ))$ となるべアμ, μ'(∈Pc')と$\pi^\circ \in \Pi^\circ$ がみつかったならば、$E_i(f(\mu, \pi^\circ)) \oplus E_i(f(\mu', \pi^\circ))$ からある付随語νを選びν_{i+1} を$\pi^\circ\nu$としてステップ i+1 に行く．そうでなければm=i として有限状態受理機M を構成する

$M = <K, \Pi^\circ, \delta, q_0, F>$において、$K = \{E_m(\mu) \mid \mu \in Tc\}$、$q_0 = E_m(\lambda)$、$F = \{E_m(\mu) \mid E_m(\mu) \ni \lambda\}$ であり、δは次のように定義される： $E_m(\mu)$が空ならば、すべての$\pi^\circ \in \Pi^\circ$に対して$E_m(\mu)$自身に遷移する。そうでないならば、$E_m(\mu)=E_m(\mu')$となるμ'∈Pcが存在して、π°に対して状態$E_m(\mu'\pi^\circ)$に遷移する。

アルゴリズムLID はアルファベットΣの個数とPcの要素の個数の多項式個の質問でM を同定する。LID の正当性は文献[3]を参照。

謝辞 国際研・学習システム研究グループの諸氏に感謝する。なお、本研究は第五世代コンピュータプロジェクトの一環として行ったものである。

参考文献

[1] Angluin,D.(1981),"A Note on the number of Queries Needed to Identify Regular Languages", Information and Control,51,76-87.

[2] Ginsburg,S. and Spanier,E.H.(1968),"Control Sets on Grammars",Math. Systems Theory,2,159-177.

[3] Takada,Y.(1987),"A Constructive Method of Grammatical Inference for Linear Languages based on Control Sets", in preparation.