数理解析研究所講究録
1041 巻 1998 年 151-158

誤情報を含む正則パターン言語の多項式時間推論

竹内　正幸*　　佐藤　優子†
Masayuki Takeuchi　Masako Sato

*大阪府立大学大学院　総合科学研究科
†大阪府立大学　総合科学部

〒 599-8531 大阪府堺市学園町 1–1

要旨. 本稿では, 誤情報を含む例から目標言語を推論する枠組みを提案する. この推論の枠組みでは, 目標言語に含まれる語が全て提示され, 目標言語に含まれない誤った語 (誤情報) も提示される. 目標言語の誤情報を含む提示を定式化するため, 近傍系と呼ばれる概念を導入する. 言語 L の近傍系とは, 言語 L と言語 L を包含する言語から成る集合で, この近傍系に属する言語の正提示を言語 L の誤情報を含む提示とする. 推論機械が目標言語 L の近傍系の任意の言語 L' と L' の任意の正提示に対して, L を極限同定するとき, 推論機械はその近傍系に関する誤情報を含む例から目標言語 L を極限同定するという. 帰納的言語の添え字付き族に対して, 各言語の近傍系を定義し, その近傍系に関する誤情報を含む例からの帰納推論, 無矛盾性, 保守性, 多項式時間で更新する推論機械等の概念を導入する. また, 上記の帰納推論を具体的に実現するため, 正則パターン言語を採用し, 正則パターン言語の族に対して, 誤情報を含む例から多項式時間の更新で無矛盾かつ保守的に推論可能となる近傍系を与える.

1. はじめに

帰納推論とは, 具体的な事例からそれらの例を説明する一般的な概念 (規則) を推論する過程である. 言語の推論の場合, 一般的な概念とは文法やオートマトン等, 言語を生成, 或いは, 認識する機構が考えられる.「極限同定」と呼ばれる帰納推論の成功基準に基づく枠組みが, Gold [6] によって提案され, その後, 言語や帰納的関数の推論問題が数多く研究されてきた. 特に, 言語の帰納推論では, 正提示と呼ばれる目標言語に含まれる事例の提示による推論問題が活発に研究されている. 言語 L の正提示とは, 言語 L に含まれる語の無限列で, 言語 L に含まれる語は少なくとも 1 回は出現する列である. 従って, 言語 L に含まれない誤った例 (誤情報) は許されていない. しかし, 実際的な場面で与えられる例は必ずしも, 目標言語に含まれる例だけとは限らない. そして, このような誤情報は, 目標言語と無関係な事例ではなく, 目標言語になんらかの意味で「近い事例」ではないだろうか. このような観点から言語の誤情報を含む提示を定式化するため, 近傍系と呼ばれる概念を導入する. 言語の近傍系とは, 言語とその言語を包含する言語から成る集合のことをいう. 言語 L に対して, 言語 L だけから成る集合 $\{L\}$ も言語 L の近傍系であり, 言語 L に有限個の語を加えた言語の集合 $\{L \cup S \mid S$ は有限言語である.$\}$ も言語 L の近傍系である.

言語 L の近傍系 $\mathcal{N}_L$ が定義されているとき, 言語 L の近傍系 $\mathcal{N}_L$ に関する誤情報を含む提示とは, $\{w_n \mid n \geq 1\} \in \mathcal{N}_L$ を満たす語の無限列 $w_1, w_2, \cdots$ のことをいう. 明らかに, この提示は言語

L の語を全て含んでいる. また, 言語 L の近傍系がその言語 L だけから成る場合, 目標言語の近傍系に関する誤情報を含む提示とは, 言語 L の正提示に他ならない.

本稿で扱う誤情報を含む例からの帰納推論も「極限同定」の枠組みを採用するが, 目標言語に関する例の提示は上で述べた誤情報を含む提示とする. 誤情報を含む帰納推論に関して, Stephan [13], Case et al. [5] 等の Noisy 推論の研究がある. そこで導入された目標言語 L に対する提示 (Noisy text) とは, 言語 L に含まれる語は無限回出現し, 含まれない語は有限回しか出現しない語の無限列である. その提示に出現する語の集合は, 言語 L と有限言語の和であるので, 本稿の枠組みで考えると上で述べた言語 L の近傍系 $\{L \cup S \mid S$ は有限言語である.$\}$ として考えることができる.

言語 L の近傍系 $\mathcal{N}_L$ が定義されているとする. このとき, 推論機械 M が言語 L の近傍系 $\mathcal{N}_L$ の任意の言語とその言語の任意の正提示に対して, 言語 L を極限同定するとき, 推論機械 M は言語 L をその近傍系 $\mathcal{N}_L$ に関する誤情報を含む例から極限同定するという. $\mathcal{L} = L_1, L_2, \cdots$ を帰納的言語の添え字付き族とし, 族 $\mathcal{L}$ の各言語 L_i にはその近傍系 $\mathcal{N}_i$ が定義されているとする. 推論機械 M が任意の添え字 i に対して, 言語 L_i をその近傍系 $\mathcal{N}_i$ に関する誤情報を含む例から極限同定するとき, 推論機械 M が言語族 $\mathcal{L}$ をその近傍系に関する誤情報を含む例から推論するという. このような定式化の下では, $L \in \mathcal{N}_i \cap \mathcal{N}_j$ $(i \neq j)$ を満たす言語 L が存在すれば, どんな推論機械を用いてもこの言語 L の正提示が入力されたとき, 言語 L_i と言語 L_j の双方を極限同定することはできない. 従って, 各言語の近傍系が互いに素であること (このような族の近傍系を無矛盾という) は, 言語族がその近傍系に関する誤情報を含む例から推論可能であるための必要条件である. また, 正の例からの帰納推論と同様に, 言語 (族) の近傍系における推論の無矛盾性, 保守性, 多項式時間で更新する推論機械等の概念を導入する.

本稿では, 上記の誤情報を含む例からの帰納推論の枠組みを具体的に提示するため, 正則パターン言語を採用する. そして, (1) 目標言語に何らかの意味で近い近傍系, (2) 効率的な推論アルゴリズムが構築できる近傍系, という観点から Hamming 距離を用いて各正則パターン言語の近傍系を構成する. パターン言語は, Smullyan [12] によって導入された基本形式体系 (Elementary Formal System) の最下位に位置する基本的かつ重要な言語であり, また実際的な応用面でも, ゲノム情報等の分野で関心がもたれ, 多くの注目を集めている. 正則パターン言語に関する性質や推論可能性に関しては, Angluin [1], Angluin [2], Arimura et al. [4], Shinohara et al. [10], Shinohara [11], Wright [14] を参照されたい. なお, 紙面の都合により定理などの証明は全て省略する.

2. 誤情報を含む例からの帰納推論

この節では, 誤情報を含む例からの帰納推論という新しい枠組みを提案する.

Σ をアルファベットとする. Σ 上の全ての語 (有限な文字列) の集合を Σ^* で表し, 空語を除いた全ての語の集合を Σ^+ で表す. Σ^* の部分集合を言語という. 自然数の集合を N で表す.

言語 L の正提示とは, $\{w_i \mid i \in N\} = L$ を満たす語の無限列 $w_1, w_2, \cdots$ のことをいう. 語の無限列 σ の n 番目までの有限初期部分列を $\sigma[n]$ で表し, $\sigma[n]$ に出現する語の集合を $\hat{\sigma}[n]$ で表す.

言語 L の近傍系 $\mathcal{N}_L$ とは, $L \in \mathcal{N}_L$ でかつ, 任意の $L' \in \mathcal{N}_L$ が言語 L を包含する言語から成る集合のことをいう. 言語 L だけから成る集合 $\{L\}$ も言語 L の近傍系であり, 言語 L に有限個の語を加えた言語の集合 $\{L \cup S \mid S$ は有限言語である.$\}$ も言語 L の近傍系である.

言語 L の近傍系 $\mathcal{N}_L$ が定義されているとする. このとき, 言語 L の近傍系 $\mathcal{N}_L$ に関する誤情報を含む提示とは, $\{w_n \mid n \geq 1\} \in \mathcal{N}_L$ を満たす語の無限列 $w_1, w_2, \cdots$ のことをいう. 明らかに, この提示には言語 L の語が全て出現している. また, 言語 L の近傍系が言語 L だけから成る場合, 言語 L の近傍系 $\mathcal{N}_L$ に関する誤情報を含む提示とは, 言語 L の正提示に他ならない.

言語族 $\mathcal{L} = L_1, L_2, \cdots$ が帰納的言語の添え字付き族であるとは, (i) $w \in L_i$ のとき, $f(w,i) = 1$, (ii) $w \notin L_i$ のとき, $f(w,i) = 0$ を満たす帰納的関数 $f : \Sigma^* \times N \to \{0,1\}$ が存在することをいう. 言語 L_i の添え字 i は, L_i を生成, 或いは, 認識するパターンやオートマトン等の記述を意味する. 以下, 言語族は帰納的言語の添え字付き族とし, 空言語は考えないものとする.

推論機械 M とは, ときどき入力を要求して, ときどき出力を生成する有効な手続きのことをいう. 推論機械によって生成する出力を推測という. 語の無限列 σ に対して, その有限列 $\sigma[n]$ が入力された後, 推論機械 M が生成する出力を $M(\sigma[n])$ で表す. 推論機械 M が語の無限列 σ に対して, 自然数 i に収束するとは, ある $m \geq 1$ が存在し, 任意の $n \geq m$ に対して, $M(\sigma[n]) = i$ となることをいう. 推論機械 M が言語 L をその近傍系 $\mathcal{N}_L$ に関する誤情報を含む例から極限同定するとは, 言語 L の近傍系 $\mathcal{N}_L$ に関する任意の誤情報を含む提示に対して, $L = L_i$ となる自然数 i に収束することをいう.

言語族 $\mathcal{L} = L_1, L_2, \cdots$ の各言語 L_i にはその近傍系 $\mathcal{N}_i$ が定義されているとする. このとき, $\{\mathcal{N}_i \mid i \in N\}$ を言語族 $\mathcal{L}$ の近傍系という.

推論機械 M が言語族 $\mathcal{L}$ をその近傍系に関する誤情報を含む例から推論するとは, 推論機械 M が任意の添え字 i に対して, 言語 L_i をその近傍系 $\mathcal{N}_i$ に関する誤情報を含む例から極限同定することをいう. 言語族 $\mathcal{L}$ がその近傍系に関する誤情報を含む例から推論可能であるとは, 族 $\mathcal{L}$ をその近傍系に関する誤情報を含む例から推論する推論機械が存在することをいう.

言語族 $\mathcal{L}$ の近傍系が無矛盾であるとは, 各 L_i の近傍系 $\mathcal{N}_i$ が互いに素, 即ち, $\mathcal{N}_i \cap \mathcal{N}_j = \phi$ $(i \neq j)$ であることをいう. 明らかに, 言語族の近傍系が無矛盾でないならば, 誤情報を含む例から推論可能ではない. 実際, $L \in \mathcal{N}_i \cap \mathcal{N}_j$ $(i \neq j)$ を満たす言語 L が存在すれば, どんな推論機械を用いてもこの言語 L の正提示が入力されたとき, 言語 L_i と言語 L_j の双方を極限同定することはできないからである.

言語 L の近傍系 $\mathcal{N}_L$ が語の有限集合 S に対して無矛盾であるとは, $S \subseteq L'$ を満たす $L' \in \mathcal{N}_L$ が存在することをいう.

推論機械 M が言語族 $\mathcal{L}$ の近傍系 $\{\mathcal{N}_i \mid i \in N\}$ に関して無矛盾であるとは, 任意の i と近傍系 $\mathcal{N}_i$ の任意の提示 σ に対して, n 時点における推論機械 M の推測を $g_n (= M(\sigma[n])$ とするとき, L_{g_n} の近傍系 $\mathcal{N}_{g_n}$ が集合 $\hat{\sigma}[n]$ に対して無矛盾であることをいう. 推論機械 M が保守的であるとは, 言語 $L_{g_{n-1}}$ の近傍系 $\mathcal{N}_{g_{n-1}}$ が集合 $\hat{\sigma}[n]$ に対して無矛盾ならば, 推論機械は出力を変更しない, 即ち, $g_n = g_{n-1}$ であることをいう.

言語族 $\mathcal{L}$ がその近傍系に関する誤情報を含む例から多項式時間の更新時間で無矛盾かつ保守的に推論可能であるとは, 族 $\mathcal{L}$ をその近傍系に関する誤情報を含む例から族 $\mathcal{L}$ の近傍系に関して無矛盾かつ保守的に推論する推論機械 M が存在し, n 番目の入力を受け取った後, 推測 g_n を出力として生成するまでの時間がそれまでの入力の長さの和に関する多項式で計算することをいう.

3. 正則パターン言語

この節では, 誤情報を含む例からの推論の枠組みを具体的に提示するため, Shinohara [10] によって導入された「正則パターン言語」に関する定義を与える. 正則パターンは, Angluin [1] によって定義された「パターン」に変数の出現回数の制限を付加したものである.

Σ を少なくとも 2 個の定数を含むものとする. Σ と互いに素な可算集合を X で表す. Σ, X の要素をそれぞれ定数（記号）, 変数（記号）という.

パターンとは, $\Sigma \cup X$ 上の空でない有限な文字列のことである. 同じ変数が高々 1 個出現するパターンのことを正則パターンという. 全ての正則パターンの集合を $\mathcal{RP}$ で表す. パターン p の長さ

を $|p|$ で表す. パターン p の位置 i の記号を $p[i]$ で表す. 2 つのパターン p,q の連接を pq で表す. パターンの集合 P に対して, 集合 P の個数を $\sharp P$ で表す. パターンの集合 P と任意の $p \in P$ に対して, 集合 $\{p' \in P \mid p' \neq p\}$ を $P \setminus p$ で表す.

代入とは, 全ての定数をそれ自身に写すパターンからパターンへの連接を保存する準同型写像のことをいう. 代入 θ によるパターン p の像を $p\theta$ で表す. 以下, 代入として, 非消去代入, 即ち, $|x\theta| \geq 1\ (x \in X)$ を満たす代入 θ を考えるものとする. 変数の付け替えとは, 任意の変数 x, y に対して, $x\theta \in X$ で, $x \neq y$ ならば, $x\theta \neq y\theta$ を満たす代入 θ のことをいう.

パターン p, q が等価であるとは, $p = q\theta$ を満たす変数を付け替える代入 θ が存在することをいう. パターン q が p の汎化であるとは, $p = q\theta$ を満たす代入 θ が存在することをいい, $p \preceq q$ で表す. 特に, $p \preceq q, q \not\preceq p$ を満たすとき, $p \prec q$ で表す. 明らかに, $p \preceq q$ ならば, $|p| \geq |q|$ である. また, p, q が等価であることと, $p \preceq q, q \preceq p$ であることは同値である. 従って, 等価なパターンを同一視すると, 関係 $\preceq$ は半順序関係である.

パターン p とパターン集合 P に対して, $P \sqsubseteq \{p\}$ であるとき, p は集合 P の汎化という. 特に, $p' \prec p$ となる P の汎化 p' が存在しないとき, この汎化 p を P の極小汎化といい, P の極小汎化の中で長さが最も長いパターンを P の最長極小汎化という.

パターン p の生成する言語とは, 集合 $\{w \in \Sigma^+ \mid w \preceq p\}$ のことをいい, $L(p)$ で表す. 言語 L がパターン言語であるとは, $L(p) = L$ を満たすパターンが存在することをいう. パターン p, q に対して, $p \preceq q$ ならば, $L(p) \subseteq L(q)$ であり, $L(p) \subseteq L(q)$ ならば, $|p| \geq |q|$ である. 正則パターン言語の族を $\mathcal{RPL}$ で表す. 言語 $L(p)$ の最短長の語から成る集合を $S_1(p)$ で表す. パターンの集合 P に対して, $L(P) = \bigcup_{p \in P} L(p)$, $S_1(P) = \bigcup_{p \in P} S_1(p)$ とする.

4. 正則パターン言語の和で構成される近傍系

この節では, 正則パターン p に対して, 近傍正則パターン集合の定義を与え, この集合の正則パターンで生成される言語の和を構成要素とする $L(p)$ の近傍系について論じる. 無矛盾となる具体的な族 $\mathcal{RPL}$ の近傍系を提示し, 正則パターン言語の族がその近傍系に関する誤情報を含む例から多項式時間の更新で無矛盾かつ保守的に推論可能であることを示す.

正則パターン p に出現する定数の位置を示す添え字の集合を $I_c(p) = \{i \mid p[i] \in \Sigma\}$ で表す.

p, q を $I_c(p) = I_c(q)$ を満たす等長な正則パターンとする. このとき, パターン p とパターン q の距離を $p[i] \neq q[i]$ を満たす $i \in I_c(p)$ の個数とし, $d(p,q)$ で表す. 正則パターン p に対して, $\mathcal{RP}_{p\langle k\rangle} = \{q \in \mathcal{RP} \mid d(p,q) \leq k\}$ とおく. 特に, 集合 $\mathcal{RP}_{p\langle 1\rangle}$ を $\mathcal{RP}_p$ で表す. 集合 $\mathcal{RP}_{p\langle k\rangle}$ の部分集合で, 正則パターン p を含む集合 P をパターン p の k-近傍正則パターン集合といい, パターン p をその集合の核パターンという. $K = \sharp\Sigma$ とすると, k-近傍正則パターン集合 P は高々 $K^{|p|}$ 個の正則パターンから成る集合である. 正則パターン p のすべての k-正則パターン集合の族を k-$\mathcal{NRP}_p$ で表し, k-近傍正則パターン集合 P から生成される言語 $L(P)$ からなる族を k-$\mathcal{NRPL}_p$ で表す. 特に, 1-近傍正則パターン集合を単に近傍パターン集合といい, 族 1-$\mathcal{NRP}_p$, 1-$\mathcal{NRPL}_p$ をそれぞれ, $\mathcal{NRP}_p$, $\mathcal{NRPL}_p$ で表す. 族 $\{k\text{-}\mathcal{NRPL}_p \mid p \in \mathcal{RP}\}$ は族 $\mathcal{RPL}$ の近傍系ではあるが, 無矛盾ではない. 実際, 次の例からこのことが示される.

例 4.1. $p = axbc$, $q = bybc$ とし, $P = \{axbc, bxbc\}$ とすると, P は p, q の近傍パターン集合である. 故に, $L(P) \in \mathcal{NRPL}_p \cap \mathcal{NRPL}_q$ である. 従って, 族 $\mathcal{RPL}$ の近傍系 $\{\mathcal{NRPL}_p \mid p \in \mathcal{RP}\}$ は無矛盾ではない. これは 族 $\mathcal{RPL}$ の近傍系 $\{k\text{-}\mathcal{NRPL}_p \mid p \in \mathcal{RP}\}$ が無矛盾ではないことを意味する.

従って, 族 $\mathcal{RPL}$ はその近傍系 $\{k\text{-}\mathcal{NRL}_p \mid p \in \mathcal{RP}\}$ に関する誤情報を含む例から推論可能で

ない．そこで，$k=1$ の場合に着目し，族 $\mathcal{NRP}_p$ の部分族で次のような条件を満たす近傍系を考える．

定義 4.2. 正則パターン p と p の近傍パターン集合 P に対して，次の条件を定義する:

条件 A_1 : 任意の $i \in I_c(p)$ に対して，$\{p'[i] \in \Sigma \mid p' \in P\} \subsetneq \Sigma$ である．

条件 A_2 : $\sharp P \geq 2$ ならば，任意の $i \in I_c(p)$ に対して，$P \not\subseteq \{p_{i,a} \mid a \in \Sigma\}$ である．
但し，$p_{i,a}$ は p の位置 i の記号を $a \in \Sigma$ で置き換えた正則パターンを表す．

条件 A : P は条件 A_1 と条件 A_2 を共に満たす．

条件 A_i $(i=1,2)$ (resp. A) を満たす近傍パターン集合を A_i (resp. A) 近傍パターン集合という．p の A_i (resp. A) 近傍パターン集合の全体から成る族を $\mathcal{NRP}_p^{A_i}$ (resp. $\mathcal{NRP}_p^{A}$) で表し，この族から生成される言語の族を $\mathcal{NRPL}_p^{A_i}$ (resp. $\mathcal{NRPL}_p^{A}$) で表す．即ち，正則パターン p に対して，

$$\mathcal{NRPL}_p^{A_i} = \{L(P) \mid P \in \mathcal{NRP}_p^{A_i}\}, \quad i=1,2, \qquad \mathcal{NRPL}_p^{A} = \{L(P) \mid P \in \mathcal{NRP}_p^{A}\}$$

である．条件 A_1 を満たす近傍パターン集合の性質として次の結果が得られる．

定理 4.3. p, q を正則パターンとし，P, Q をそれぞれ p, q の A_1 近傍パターン集合とする．このとき，次の (1)(2)(3) は同値である．

(1) $L(P) = L(Q)$. (2) $S_1(P) = S_1(Q)$. (3) $P = Q$.

また，条件 A_1 の定義から明らかに，$\sharp\Sigma = 2$ ならば，正則パターン p の A_1 近傍正則集合は $\{p\}$ だけである．従って，$\mathcal{NRPL}_p^{A_1} = \{L(p)\}$ である．一方，$\sharp\Sigma \geq 3$ ならば，族 $\mathcal{RPL}$ の近傍系 $\{\mathcal{NRPL}_p^{A_1} \mid p \in \mathcal{RP}\}$ は無矛盾ではない．実際，例 4.1で与えたパターン集合 $P = \{axbc, bxbc\}$ は，$p = axbc$ と $q = bxbc$ の 2 つの異なる正則パターン (核パターン) に対して条件 A_1 を満たす近傍パターン集合である．従って，条件 A_1 を満たす近傍パターン集合では，必ずしも核パターンは一意には定まらない．しかし，条件 A_1 を満たす上記の近傍系に対して，本稿で与える推論手続きは必ずしも正しい核パターンには収束するとは限らないが，推論過程で生成される A_1 近傍パターン集合は入力された提示に対応する正しい P に収束する．

A_1 近傍パターン集合の全体の族を $\mathcal{NRP}^{A_1} = \bigcup_{p \in \mathcal{RP}} \mathcal{NRP}_p^{A_1}$ とおく．近傍パターン集合 P が語の集合 S の 族 $\mathcal{NRP}^{A_1}$ における極小多重汎化であるとは，$S \subseteq L(P)$ でかつ，$S \subseteq L(Q) \subsetneq L(P)$ を満たす $Q \in \mathcal{NRP}^{A_1}$ が存在しないことをいう．A 近傍パターン集合の族に対しても同様に定義する．このとき，上記の定理から直ちに次の定理が成立する．

定理 4.4. 任意の $P \in \mathcal{NRP}^{A_1}$ は，族 $\mathcal{NRP}^{A_1}$ における集合 $S_1(P)$ の極小多重汎化である．

補題 4.5. S を語の集合とする．このとき，次の手続き MMG は，集合 S の族 $\mathcal{NRP}^{A_1}$ における極小多重汎化 P とその核パターン p を集合 S に含まれる語の長さの総和に関する多項式時間で計算する．

Procedure MMG
input : a finite set S of words ;
output : a regular pattern p and a neighborhood set P of p ;
begin
compute the longest minimal generalization of (S) and let p be its output ;
let $P := \phi$, $K := \sharp\Sigma$, $m := |p|$, $t := 0$;
if there are i and j such that $S \subseteq \bigcup_{a \in U_j} L(p_{i,a})$, $p[i] \in X$ and $j < 2^K - 1$ **then**

let $i_1 := min\{i \mid S \subseteq \bigcup_{a \in U_j} L(p_{i,a}), p[i] \in X\}$;
let $j_1 := min\{j \mid S \subseteq \bigcup_{a \in U_j} L(p_{i_1,a})\}$, $P := \{p_{i_1,a} \mid a \in U_{j_1}\}$;
for each $p \in P$ **do if** $t = 0$ **then begin**
let $P := P \setminus p$;
for $i_2 = i_1 + 1$ **to** m **do if** $p[i_2] \in X$ **then**
if there are j and b such that $S \subseteq \bigcup_{b' \in U_j} L(p_{i_2,b'}) \cup L(P_{i_2,b})$ **then begin**
let $j_2 := min\{j \mid S \subseteq \bigcup_{b' \in U_j} L(p_{i_2,b'}) \cup L(P_{i_2,b})\}$;
let b be a constant such that $S \subseteq \bigcup_{b' \in U_{j_2}} L(p_{i_2,b'}) \cup L(P_{i_2,b})\}$;
let $P := P_{i_2,b} \cup \{p_{i_2,b'} \mid b' \in U_{j_2} \setminus b\}$, $p := p_{i_2,b}$, $t := 1$;
end ;
end ;
let $P := P \cup \{p\}$;
output p and P ;
end .

但し, U_j は $\Sigma = \{a_1, \cdots, a_n\}$ の部分集合を空集合を除いて, 集合の個数の小さい順に枚挙した列

$$\{a_1\}, \cdots, \{a_n\}, \{a_1, a_2\}, \cdots, \{a_{n-1}, a_n\}, \{a_1, a_2, a_3\}, \cdots, \{a_1, \cdots, a_n\}$$

の j 番目の集合を表し, $P_{i,a}$ はパターンの集合 P と位置 i, 及び, 定数 a に対して, $\{p_{i,a} \mid p \in P\}$ を表す. また, 語の集合 S の最長極小汎化を求める手続きと正則パターン言語の所属問題に関する結果は, Shinohara [10] を参照されたい.

上記の条件 A_2 は, 後述するように無矛盾性を保証する条件である. この条件は, $\sharp P \geq 2$ ならば, $\sharp P \geq 3$ を意味する. 実際, $\sharp P \geq 2$ ならば, P は少なくとも 2 つの $i, j \in I_c(p)$ $(i \neq j)$ に対して, p と異なる正則パターン $p_{i,a}$, $p_{j,b}$ $(a, b \in \Sigma)$ を含むことを要求しているからである.

補題 4.6. p を正則パターンとし, P を p の A_2 近傍パターン集合とする. このとき, P は p 以外の正則パターンの A_2 近傍パターン集合とはならない.

定理 4.7. 族 $\mathcal{RPL}$ の近傍系 $\{\mathcal{NRPL}_p^{A_2} \mid p \in \mathcal{RP}\}$ は無矛盾である.

ここで, 条件 A を満たす近傍パターン集合について考える. 条件 A の定義から明らかに, $\mathcal{NRPL}_p^A = \mathcal{NRPL}_p^{A_1} \cap \mathcal{NRPL}_p^{A_2}$ となり, 上で得られた結果は全て条件 A の下でも成立することが示される.

例 4.8. $\Sigma = \{a, b, c\}$ とし, $p_1 = axa$, $p_2 = axy$, $p_3 = aaa$ とすると, 族 $\mathcal{NRP}_{p_1}^A, \mathcal{NRP}_{p_2}^A, \mathcal{NRP}_{p_3}^A$ は次のように与えられる.

$$\mathcal{NRP}_{p_1}^A = \{\{axa\}, \{axa, bxa, axb\}, \{axa, bxa, axc\}, \{axa, cxa, axb\}, \{axa, cxa, axc\}\},$$
$$\mathcal{NRP}_{p_2}^A = \{\{axy\}\},$$
$$\mathcal{NRP}_{p_3}^A = \{\{aaa\}, \{aaa, aab, aba\}, \{aaa, aac, aba\}, \cdots, \{aaa, aac, \cdots, caa\}\}.$$

以上の結果をまとめると, 次の結果が示させる.

定理 4.9. 族 $\mathcal{RPL}$ はその近傍系 $\{\mathcal{NRPL}_p^A \mid p \in \mathcal{RP}\}$ に関する誤情報を含む例から多項式時間の更新で無矛盾かつ保守的に推論可能である.

5. 一般の言語で構成される近傍系

この節では, 各正則パターン言語に対し, 必ずしも正則パターン言語の和とはならない一般的な言語を構成要素とする近傍系を定義する. ここで導入する近傍系は, 前節で提案した近傍系を含む近傍系である. そして, 正則パターン言語の族がその近傍系に関する誤情報を含む例から多項式時間の更新で無矛盾かつ保守的に推論可能であることを示す.

w, w' を等長な語とする. このとき, w と w' に対して, $w[i] \neq w[i]$ を満たす i の個数 (Hamming 距離) を $d(w, w')$ で表す. 言語 L' が言語 L の k-近傍言語であるとは, $L \subseteq L'$ で, 任意の $w' \in L'$ に対して, $d(w, w') \leq k$ を満たす $w \in L$ が存在することをいう.

定理 5.1. p を正則パターン, L を $L(p)$ の k-近傍言語とする. このとき, $L \subseteq L(P)$ を満たす p の k-近傍正則パターン集合 P が存在する.

以下, 前節と同様に $k = 1$ の場合だけを扱う. 正則パターン言語 $L(p)$ の 1-近傍言語を単に近傍言語といい, 族 1-$\mathcal{NL}_p$を $\mathcal{NL}_p$で表す. また, p を族 $\mathcal{NL}_p$の核パターンという.

定理 5.2. 任意の正則パターン p に対して, $\mathcal{NRPL}_p \subseteq \mathcal{NL}_p$ である. 従って, $\{\mathcal{NL}_p \mid p \in \mathcal{RP}\}$ は正則パターンの族 $\mathcal{RPL}$ の無矛盾とはならない近傍系である.

前節と同様に, 無矛盾な近傍系を得るために次の条件を満たす近傍言語から構成される族 $\mathcal{NL}_p$ の部分族を定義する.

定義 5.3. p を正則パターンとする. このとき, 言語 $L(p)$ の近傍言語 L が条件 B を満たすとは,

1. $L \subseteq L(P)$ を満たす A 近傍パターン集合 P が存在する.
2. $L(p) \subsetneq L$ ならば, $w_i \in L(p_i) - L(\mathcal{RP}_{p\langle 2\rangle} \backslash p_i)$ $(i = 1, 2)$ を満たす $w_1, w_2 \in L$ と $p_1, p_2 \in \mathcal{RP}_p$ $(p_1, p_2 \neq p,\ p_1 \neq p_2)$ が存在する.

を満たすことをいう.

条件 B を満たす正則パターン言語 $L(p)$ の近傍言語 L を B 近傍言語といい, B 近傍言語から成る族を $\mathcal{NL}_p^B$ で表す. 即ち, 正則パターン言語 $L(p)$ に対して,

$$\mathcal{NL}_p^B = \{L \subseteq \Sigma^+ \mid L \text{ は } L(p) \text{ の B 近傍言語である.}\}.$$

である. 明らかに, 集合 $\{\mathcal{NL}_p^B \mid p \in \mathcal{RP}\}$ は族 $\mathcal{RPL}$ の近傍系である. 前節で扱った近傍系 $\mathcal{NRPL}_p^A$ は有限個の言語 (正則パターン言語の和) であるが, この節で導入した $\mathcal{NL}_p^B$ は無限個の言語から構成される近傍系である. これらの近傍系の間には次の結果が成り立つ.

定理 5.4. 任意の正則パターン p に対して, $\mathcal{NRPL}_p^A \subseteq \mathcal{NL}_p^B$ が成立する.

補題 5.5. p を正則パターンとし, L を $L(p)$ の B 近傍言語とする. このとき, 言語 L に対する核パターンは p だけである.

定理 5.6. 族 $\mathcal{RPL}$ の近傍系 $\{\mathcal{NL}_p^B \mid p \in \mathcal{RP}\}$ は無矛盾である.

以上の結果をまとめると, 次の結果が示される.

定理 5.7. 族 $\mathcal{RPL}$ はその近傍系 $\{\mathcal{NL}_p^B \mid p \in \mathcal{RP}\}$ に関する誤情報を含む例から多項式時間の更新で無矛盾かつ保守的に推論可能である.

6. おわりに

本稿では, 正則パターンに関する 2 種類の近傍系を提示し, その近傍系に関する誤情報を含む例からの推論を展開した. どちらの近傍系も目標言語の語と高々 1 個の定数が異なるという意味で目標言語に近いと思われるが, その近傍系として本当に近い (条件 A_2 を満たさない) 言語を許していない. もしそれを許すならば, 2 節で述べたように推論不可能になる. しかし, 条件 A_2 を満たさない近傍系に対しても, 本稿で導入した推論アルゴリズムでは, 正しいパターンか, 間違っていたとしても誤っている位置を指定できるパターンを出力できる. 条件 A_1 は推論アルゴリズムを簡単にするために導入した. この条件は, 本稿で用いた最短の語の集合 $S_1(P)$ を含むもっと大きい集合に関して定理 4.4 を示せば, 取り外すことが出来る条件である. 今後, Hamming 距離に限らず定数の消去等, 意味のある近傍系に対する誤情報を含む例からの推論の展開が課題である.

参考文献

[1] D. Angluin. "*Finding patterns common to a set of strings*", Information and Control, vol. 21, 46-62, 1980.

[2] D. Angluin. "*Inductive inference of formal languages form positive data*", Information and Contorol, vol. 45, 117-135, 1980.

[3] S. Arikawa, S. Miyano, A. Shinohara, S. Kuhara, Y. Mukouchi and T. Shinohara. "*A machine discovery from amino acid sequences by decision trees over regular patterns*", New Generation Computing, vol. 11, 361-375, 1993.

[4] H. Arimura, T. Shinohara and S. Otsuki. "*Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data*", Proc. 11th STACS, LNCS 775, 649-660, 1994.

[5] J. Case, S. Jain and A. Sharma. "*Synthesizing noise-tolerant language learners*", Proc. 8th International Workshop on Algorithmic Learning Theory, 228-243, 1997.

[6] E. M. Gold. "*Language identification in the limit*", Information and Control, vol. 10, 447-474, 1967.

[7] T. Motoki, T. Shinohara and K. Wright. "*The correct definition of finite elasticity : corrigendum to identification of unions*", Proc. 4th Workshop on Computational Learning Theory, 375-375, 1991.

[8] Y. Mukouchi. "*Containment problems for pattern languages*", IEICE Trans. Inf. & Syst., Vol. E75-D, No. 4, 420-425, July 1992.

[9] M. Sato. "*Inductive inference of formal language*", Bulletin of Informatics and Cybernetics, vol. 27, No. 1,pp. 85-106, 1995.

[10] T. Shinohara. "*Studies on inductive inference from positive data*", PhD thesis, University of Kyushu, 1986.

[11] T. Shinohara and H. Arimura. "*Inductive inference of unbounded unions of pattern languages from positive data*", Proc. 7th International Workshop on Algorithmic Learning Theory, 256-271, 1996.

[12] R. M. Smullyan. "*Theory of Formal Systems*", Prinston University Press, Princeton, New Jersey, 1961.

[13] F. Stephan. "*Noisy Inferencee and Oracles*", Proc. 6th International Workshop on Algorithmic Learning Theory, 185-200, 1995.

[14] K. Wright. "*Identification of unions of languages drawn from positive data*", Proc. 2nd Annual Workshop on Computational Learning Theory, pp. 328-333, 1989.