極小多重汎化によるパタン和推論アルゴリズムの実験的評価

4P－7　笠井　透[†]　有村博紀[†]　篠原　武[‡]

[†]九州大学大学院システム情報科学研究科　[‡]九州工業大学情報工学部

1　はじめに

形式言語の例からの学習は，機械学習分野の主要な問題のひとつである．とくに正例からの学習は，テキストデータからの特徴やパタンの発見など応用が広い．

有村ら [2] は，正例からパタン和の学習をおこなう k **極小多重汎化** (MMG) とよばれる効率的アルゴリズムを提案し，遺伝子情報データからの知識発見に応用している．正則パタン和とは，

$$\{\ *F*M*LV*L,\ *FL*V*A*,\ *LF*M*V*\ \}$$

のような任意の文字列に照合するワイルドカード $*$ を含む文字列の有限集合である．

MMG アルゴリズムは，正例を覆う言語として極小のパタン言語和を計算し，さらに，十分に多くの例が与えられる場合は，未知のパタン和を極限で正しく同定することが証明されている．

一方，実際の応用のためには，このような極限同定可能性だけでなく，比較的少数のデータに対する学習能力を解析し，誤差やノイズに対する頑健性を明確にする必要がある．　そこで本研究では，確率的に生成したデータを用いた学習実験をおこない，MMG アルゴリズムの性能を評価した．

2　正則パタン和

定数文字の集合を Σ とする．任意の文字列 $p \in (\Sigma \cup \{*\})^+$ を**正則パタン**とよび，とくに，$*$ の出現が連続せず，その数が高々 m 個であるとき，**m 変数正則パタン**とよぶ．パタン p の**言語**とは，$*$ の出現を（それぞれ異なり得る）空文字を含む定数文字列で置き換えて得られる文字列全体の集合 $L(p)$ である．正則パタンの集合 P(以後**パタン和**とよぶ) に対して，集合和 $\bigcup_{p\in P} L(p)$ でその言語を定義する．$L(P)$ と $\Sigma^+ - L(P)$ の要素を，それぞれ**正例**および**負例**という．ここでは，仮説空間として，k 個以下の m 変数正則パタンからなるパタン和全体を考える．

3　極小多重汎化アルゴリズム

MMG アルゴリズム [2] は，最も一般的な仮説 $P = \{*\}$ から始め，現在の仮説 P が正例すべてを覆い，かつパタン数が k を超えない限り，仮説中のパタン $p \in P$ をその一個以上の精密化 $q_1, \ldots, q_\mu \in \rho(p)$ で置き換えていき，仮説を段階的に具体化していく．ただし，置き換えは，文献 [3] の極大被覆戦略 (max)，極小被覆戦略 (min)，確率的戦略 ($rand$) のいずれかの戦略にもとづいて，可能な最小数のパタンでおこなう．

精密化集合 $\rho(p)$ は，パタン p 中の任意の $*$ の出現に，任意の文字 $a \in \Sigma$ と空文字 ε に関する置換え "$* := *a*$", "$* := a*$", "$* := *a$", "$* := \varepsilon$" のどれかを，ただ一度だけ適用して得られる m 変数パタン全体として定義する．MMG アルゴリズムは，構文的汎化順序 $\sqsubseteq$ [2] に関して極小なパタン和を計算する．さらに，Σ が十分大きなとき，このパタン和は正例を覆う極小言語に一致する [2].

4　実験データの生成

学習の目標となるパタン和 P_* を**未知パタン和**という．実験では，P_* を，長さが l で，出現する変数の個数が v である h 個のパタンの和とし，可能な組合せの中から等確率で生成する．

学習に用いる正負例の多重集合を，それぞれ Pos, Neg で表す．　実験では，まず，各例の長さ d を，長さの期待値が λ であるような確率分布 $Pb(\lambda, d) = (e^{-\lambda}\lambda^d)/d!$ にしたがって決定する．さらに，正例は $L(P_*)$ から，負例は $\Sigma^+ - L(P_*)$ から，長さ d の文字列を等確率で選択して生成する．

例集合 Pos, Neg に対する**分類ノイズ**とは，正例 $s \in Pos$ がある確率で負例となるノイズである．実験では，パラメタ $r \leq \sharp Pos$ に対して，Pos から r 個の正例をのぞき，r 個の負例を加える．

5　実験

表 1 のパラメタにしたがって，仮説 P_* と例集合 Pos, Neg を生成し，訓練例の数 n を変化させて学

Empirical Evaluation of an Inference Algorithm for Unions of Patterns by Minimal Multiple Generalization

[†]Department of Informatics, Kyushu University, Fukuoka 812-8581, Japan

[‡]Department of Artificial Intelligence, Kyushu Institute of Technology, Iizuka 820-0067, Japan

表 1: データ生成パラメタと学習パラメタ

パラメタ	意味
$s = 5$	定数記号の数
$h = 1, 3, 5, 8$	未知パタン数
$l = 10$	未知パタン長
$v = 3$	未知パタン中の変数の個数
$N = 2000$	例集合 Pos,Neg の各例数
$\lambda = 5, 8, 10$	例の長さの期待値
$r = 0, 20$	分類ノイズの個数
$k = 1, 3, 5, 8$	仮説パタン数
$m = 3$	仮説パタン中の変数の個数
$e = 0, 2$	排除する例外の上限
$S \in \{max, , min, rand\}$	探索戦略
$n = 10 \sim 300$	訓練例の個数

習実験をした．本稿では，とくに指定しない限り，$h = k = 3, N = 2000, \lambda = 10$ とする．

一回の試行では，まず，Pos から訓練例を n 個ランダムに選び，指定された戦略にもとづいて MMG アルゴリズムで仮説 P を生成した．仮説 P の精度は，正負例の精度 pos と neg の相乗平均 $acc = \sqrt{pos \cdot neg}$ で評価した．ここに，$pos = \sharp(Pos \cap L(P))/\sharp Pos$, $neg = \sharp(Neg - L(P))/\sharp Neg$ である．

以上の試行を，異なる 5 つの未知パタン和に対して，それぞれ 20 回づつ繰り返し，合計 100 回の試行について精度 acc の平均と標準偏差を求めた．実験の結果は，訓練例の数 n に対する仮説の平均精度を示すグラフ（**学習曲線** とよぶ）で表した．

実験 1: ノイズの有無による仮説精度の比較実験をした．ノイズとして $r = 20$ の分類ノイズを用い，$h = k = 3, S \in min$ とし，訓練例数を $n = 10 \sim 300$ と変化させた．図 1 に学習曲線を示す．仮説精度の標準偏差は以下の表のとおりである．

表 2: 仮説精度の標準偏差

訓練例数 n	50	100	150	200	300
ノイズ無しの標準偏差	5.27	2.67	1.10	1.13	1.00
ノイズ有りの標準偏差	6.83	7.27	10.16	11.69	14.92

ノイズ無しの場合，訓練例数に対して精度が単調に増加し，学習曲線は高原状になった．一方，ノイズ有りの場合は，[2] の学習曲線に似たピークのある右肩下がりの形になった．精度を最大にする訓練例数 n が存在することがわかる．

実験 2: パタン数 $h = k$ を 1 ～ 8 と変化させて実験した．ノイズは加えず，max 戦略を用いて，$n = 10 \sim 200$ と変化させた．図 2 に，学習曲線を示す．理想的な $h = k$ の場合でも，未知パタン数 h が大きくなると例数を増やしても真の解への収束が

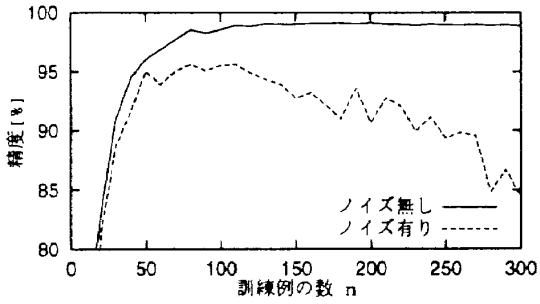

図 1: ノイズの有無による仮説精度の比較

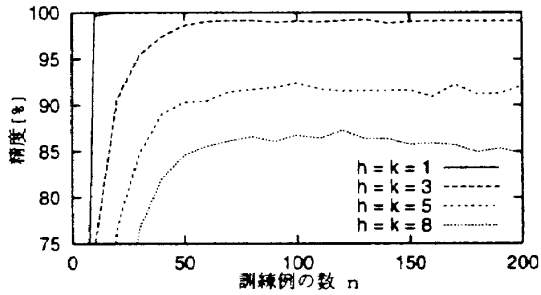

図 2: 未知パタン数による仮説精度の比較

遅く，h の値に関連した一定の精度に落ちつくことがわかる．

6 おわりに

本研究では，MMG アルゴリズムを，確率的に生成したデータからの学習実験により評価した．その結果，ノイズ有りの実験で得られた学習曲線は，遺伝子情報データによる実験 [2, 3] のものに形が似ていることがわかった．さらに，未知パタンの数が多くなるほど，学習が難しいことが確認された．

パタン和の推論について，従来の PAC 学習の枠組みは，今回調べたような実際的性能の解析には十分でないと考えられる．今後は平均値解析等の枠組みを用いて，従来の PAC 学習理論から予測できないこの種の学習過程について，より詳細な理論を構築したい．理論からの適正な訓練例数 n の予測も，今後の課題である．

参考文献

[1] D.Angluin. Inductive inference of formal language from positive data. *Inf. & Cont.* **45**, 117-135, 1980.

[2] H.Arimura, R.Fujino, T.Shinohara, and S.Arikawa. Protein Motif Discovery from Positive Examples by Minimal Multiple Generalization over Regular Patterns. In *Genome Informatics Workshop*, 39-48, 1994.

[3] 山口, 篠原, 藤野, 有村, 有川. 複数文字列パターンによるアミノ酸配列からのタンパク質モティーフの発見. 情報処理学会 情報学基礎研究会, 33-40, 1995.

[4] T.Shinohara. Polynomial time inference of extended regular pattern languages. LNCS 147, 115-127, 1982.