Empirical Evaluation of a Regular Pattern Inference Algorithm by Minimal Multiple Generalization

笠井, 透
九州大学大学院システム情報科学府情報理学専攻 : 修士課程

有村, 博紀
九州大学大学院システム情報科学府情報理学専攻

篠原, 武
九州工業大学情報工学部

https://doi.org/10.15017/1498355

出版情報：九州大学大学院システム情報科学紀要. 3 (2), pp.185-190, 1998-06-22. Faculty of Information Science and Electrical Engineering, Kyushu University
バージョン：
権利関係：

九州大学大学院
システム情報科学研究科報告
第3巻 第2号 平成10年9月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.3, No.2, September 1998

極小多重汎化による正則パタン推論アルゴリズムの実験的評価

笠井 透* ・ 有村博紀** ・ 篠原 武***

Empirical Evaluation of a Regular Pattern Inference Algorithm by Minimal Multiple Generalization

Toru KASAI, Hiroki ARIMURA and Takeshi SHINOHARA

(Received June 22, 1998)

Abstract: A regular pattern is a string consisting of constant symbols and mutually distinct variables, and represents the set of the constant strings obtained by substituting possibly empty constant strings for the variables. A learning algorithm, called k-minimal multiple generalization (k-mmg), finds a minimally general collection of at most k regular patterns that explains all the positive examples. Recently, several attempts have been made at applying this algorithm to protein motif discovery and other knowledge acquisitions. In such applications, its performance is considerably influenced by a class of data and values of learning parameters. This paper empirically evaluates performance of the algorithm k-mmg on synthesized data to apply it to protein motif discovery.

Keywords: Pattern inference, Machine learning, Positive example, Classification noise, Observation noise

1. はじめに

形式言語の例からの学習は, 機械学習の分野の主要な問題のひとつである. 未知の言語に含まれる文字列を正例といい, 含まれない文字列を負例という. 実際の問題を考えたときは, 正例からの学習は自然だが, 一般には正例からの学習は, 正負例からの学習にくらべて能力が劣ることがわかっている[1]. ここでは, 正例から学習可能であることが知られているパタン言語和を取り上げる.

正則パタンとは, 定数記号 $a,b,c,...$ と変数記号 $x,y,z,...$ からなる $xaaybbzaaw$ のような文字列で, 各変数が文字列中に高々1度しか出現しないものである. 正則パタン言語とは, 未知パタン中の変数に空文字を含む定数文字列を代入して得られる文字列全体の集合である. 正則パタンの集合 Pの言語とは, P の各パタンが表す言語の集合和をいう. 与えられた正例から未知パタンを同定することを正例からのパタン推論という.

有村ら[2),3)] は, 正例からパタン和の学習をおこなう k 極小多重汎化とよばれるアルゴリズムを提案している. このアルゴリズムは, 正例を覆う言語として極小のパタン言語和を効率よくみつける. さらに, 十分に多くの例が与えられたとき, 未知のパタン和を正しく同定することが証明されている.

文献 2) では, このパタン推論アルゴリズムをアミノ酸配列データからのタンパク質モチーフ抽出に応用している. このような実際のデータへの応用では,

(1) 仮説クラスの妥当性,
(2) データ中の誤差やノイズの影響,
(3) 学習パラメタの設定法,
(4) 仮説の精度

が問題となる. そこで, 本稿ではこれらの問題について, 確率的に生成した人工データを用いた計算機実験で調べる. そのため, 仮説と例の確率的生成と学習実験をおこなうための実験システムを設計し, これを用いて実験する. 特に, 実験では学習パラメタとして,

- 仮説中の最大パタン数 k,
- サンプルした例の長さ,
- 探索の戦略,
- データ中のノイズの有無とその種類

が学習精度におよぼす影響について明らかにすることを目標とする.

2. 準 備

2.1 正則パタンとその和言語

本節では, 文献 2),3),6) にしたがって, 正則パタン和を定義する. 集合 A に対して, $\sharp A$ で A の要素数を表し, A^* で A 上の有限文字列全体を, A^+ で $A^*-\{\varepsilon\}$ を表す. ここに, ε は**空文字**である. **定数記号**の有限集合を $\Sigma=\{a,b,...\}$ とし, **変数**の可算集合を $X=\{x,y,x_1,x_2,...\}$ とする. ただし, $X\cap\Sigma=\phi$ とする.

正則パタン[6)] とは, 定数記号と変数から構成される, 各変数が高々1度しか出現しない文字列 $p\in(\Sigma\cup X)^+$

平成10年6月22日受付
* 情報理学専攻修士課程
** 情報理学専攻
*** 九州工業大学情報工学部

である. 例えば, $p = aaxabcybb$ は正則パタンである. **代入**とは, パタンからパタンへの準同型写像 θ であり, 任意の $a \in \Sigma$ に対して $\theta(a) = a$ を満たすものである. パタン p が表す**拡張言語** $L(p)$ を, p 中の変数に空文字を含む定数文字列を代入することによって得られる文字列全体と定義する. パタンの集合 P に対して, その**和言語**を $L(P) = \bigcup_{p \in P} L(p)$ と定義する. $L(P)$ と $\Sigma^+ - L(P)$, それぞれの要素を**正例**および**負例**という.

Table 1 Example of k-mmg

A set S of positive examples		A 3-mmg of S	
e_1	WLVNFIIVIMVFILFLVGLYLL	p_1	*F*M*LV*L
e_2	VALVTITLWFMAWTPYLVINCMGL	p_2	*FL*V*A*
e_3	GFLAASALGVVMITAALAGIL	p_3	*LF*M*V*
e_4	SKILGLFTLAIMIISCCGNGVVVYI		
e_5	MTIKTSIMKILFIWMMAVFWT		
e_6	IFYSIFVYYIPLFLICYSYWFIIAAVSA		
e_7	GCGSLFGCVSIWSMCMIAFDRYNVIV		

2.2 極小多重汎化

本節では, 正則パタンの部分クラスから極小多重汎化を定義する[6]. 正則パタン p は, 変数が連続して出現しないとき**標準形**という. また m 個以下の変数を含む正則パタンを**m変数正則パタン**という. $\mathcal{RP}_\varepsilon$ と $\mathcal{RP}_{\varepsilon,m}$ で, それぞれ, 標準形正則パタン全体の集合と, その部分集合のうち m変数正則パタンからなるものを表す. $\mathcal{RP}_\varepsilon$ 上の半順序関係 $\preceq$ を, $p \preceq q \Leftrightarrow$ ある代入 θに対して$p = \theta(q)$, と定義する. 関係 $p \preceq q$ が成立するとき, q は p よりも **一般的**, p は q よりも**具体的**であるという.

正則パタンのある部分クラスを $\mathcal{D} \subseteq \mathcal{RP}_\varepsilon$ とする. 正則パタンの集合 P 中に $p \prec q$ を満たすようなパタン p, q が存在しないとき, P は**既約である**という. $\mathcal{D}$ 中の高々 k 個の既約なパタンの集合全体のクラスを $\mathcal{D}^k$ で表す. $\mathcal{D}^k$ 上の半順序関係 $\sqsubseteq$ を, $P \sqsubseteq Q \Leftrightarrow$ 任意の $p \in P$ に対して $p \preceq q$ なる $q \in Q$ が存在する, と定義する. 定義より, $P \sqsubseteq Q \Rightarrow L(P) \subset L(Q)$ だが, この逆は一般に成立しない. 逆が成立するとき, クラス $\mathcal{D}^k$ は**コンパクト性**をもつという.

定義 1 クラス $\mathcal{D}$ に対して, 正例集合 S の**k極小多重汎化** (k-mmg) とは, $S \subseteq L(P)$ を満たす $P \in \mathcal{D}^k$ で, 半順序 $\sqsubseteq$ に関して極小なものをいう.

パタン数の上限が $k = 3$ のときの k極小多重汎化の例を, **Table–1**に示す. ここに, $*$ は変数を表す. 例では, S の 3-mmg $\{p_1, p_2, p_3\}$ の表す拡張言語 $L(p_1)$, $L(p_2)$, $L(p_3)$ が, それぞれ, S の部分集合 $\{e_1, e_2\}$, $\{e_3, e_6\}$, $\{e_4, e_5, e_7\}$ を覆っている.

定理 1(文献 2),3)) パタン和のクラス $\mathcal{D}^k$ がコンパクト性をもつと仮定する. このとき, 正例の k極小多重汎化を計算するアルゴリズムは, $\mathcal{D}^k$ を多項式時間仮説更新によって正例から極限同定する.

2.3 学習アルゴリズム

本節では, k極小多重汎化を多項式時間で計算する学習アルゴリズム MMG について説明する[3),4)]. **Fig.1** から **Fig.3** にアルゴリズム MMG を示す. MMG は, 最も一般的な仮説 $\{x\}$から始めて, 各パタンを具体化したり, あるいは複数のパタンへ分割したりしながら, 仮説空間を具体的な方向へと探索する.

正則パタン p に対して, p 中に現れる変数の集合を $var(p)$ とかく. $\mathcal{RP}_{\varepsilon,m}$ に対する**精密化演算子** ρ とは, パタン p を, 以下のような操作を施して得られる m変数パタン全体の集合 $\rho(p)$ に変換するものである.

- ある変数 $x \in var(p)$ を xay で置き換える. ただし, $y \in var(p), x \neq y, a \in \Sigma$ である.
- ある変数 $x \in var(p)$ を xa または axで置き換える. ただし, $a \in \Sigma$ である.
- ある変数 $x \in var(p)$ を ε で置き換える.

集合 $\rho(p)$ は, p の長さの多項式時間で計算可能であり, ρ を用いて与えられたパタンより具体的な m変数正則パタン全体を, 段階的に計算できる.

定理 2(文献 3)) k, m を正整数, S を文字列の有限集合とする. アルゴリズム MMG は, 仮説空間 $\mathcal{RP}^k_{\varepsilon,m}$ 上の S の k極小多重汎化の一つを時間 $O(\sharp\Sigma \cdot k^3 m^k l^2 n)$ で計算する. ここに, l は S 中の最長文字列の長さであり, n は S の要素数である.

MMG が生成した仮説中には, 変数をもたないパタンが含まれることがある. このような定数文字列を**例外**という. 仮説中に例外パタンが含まれるとき, それを正例から排除することで, より高い精度の仮説を発見することが期待できる. そこで, 例外パタンをみつけると, それを仮説中には含めず, 同時に, その文字列を正例から取り除くように MMG を拡張する[5)]. ただし, 排除を許す例外の数は, あらかじめ与えたパラメータ e を上限とする.

2.4 探索戦略

学習アルゴリズム MMG では, 仮説空間の探索において, つぎの 2 種類の非決定的選択点が存在する.

(Choice1) MMG(**Fig.1**) の 4 行目でどのパタン $p \in P$ を分割するか.

(Choice2) Tighten(**Fig.3**)の 2 行目でどのパタン $p \in P$ を具体化するか.

本稿では, つぎの探索戦略にもとづいて $p \in P$ を選択

MMG(k, S) /* $k \geq 1$ *and* S *is the set of positive examples.* */
1 $P :=$ **Tighten**$(\{x\}, S)$;
2 $\Delta k := k$;
3 **while** $\Delta k \geq 2$ and there exists some $p \in P$ that is Δk-divisible(*1) member p in P with respect to $\Delta S \stackrel{\text{def}}{=} S - L(P - \{p\})$(*2) **do**
4 Choose such a divisible member p in P and the corresponding ΔS; (Choice 1)
5 $\Delta P :=$ **Divide**$(p, \Delta k, \Delta S)$;
6 $\Delta P :=$ **Tighten**$(\Delta P, \Delta S)$;
7 $P := (P - \{p\}) \cup \Delta P$;
8 $\Delta k := k - \sharp P + 1$;
9 **endwhile**;
10 **return** P;

(*1) a member p in P is Δk-*divisible* with respect to ΔS if the set **Divide**$(p, \Delta k, \Delta S)$ exists.
(*2) ΔS is the set of positive examples subsumed only by p but not by other members in P.

Fig.1 Minimal multiple generalization algorithm

Divide(p, k, S) /* p *is a pattern,* $k \geq 2$ *and* S *is a set of positive examples.* */
1 Compute the set $\rho(p)$ of one-step refinements;
2 Choose a set P of at most k members in $\rho(p)$ that is reduced with respect to(*3) S;
3 **return** P;

(*3) a set P is *reduced respect to* S if $S \subseteq L(P)$ but $S \not\subseteq L(P')$ for any proper subset $P' \subset P$.

Fig.2 Algorithm for dividing a member of ΔP into its refinements

Tighten(P, T) /* P *is a patterns and* T *is a set of positive examples.* */
1 **while** for some $p \in P$, there is some r in $\rho(q)$ such that $L(r) \supseteq \Delta T \stackrel{\text{def}}{=} T - L(P - \{q\})$(*4) **do**
2 Choose such q in P and the corresponding ΔT; (Choice 2)
3 Choose a refinement r in $\rho(q)$ such that $\Delta T \subseteq L(r)$;
4 $P := (P - \{p\}) \cup \{r\}$
5 **endwhile**;
6 **return** P;

(*4) ΔT is the set of positive examples subsumed only by q but not by other members in P.

Fig.3 Algorithm for tightening a multiple generalization by refining patterns

する.

確率的戦略: 等確率にパタン $p \in P$ を選択する.

極大被覆戦略: できるだけ多く正例を被覆するパタン $p \in P$ を選択する.

極小被覆戦略: できるだけ少なく正例を被覆するパタン $p \in P$ を選択する.

どの戦略を用いても, 学習アルゴリズムは正しく k極小多重汎化を計算するが, 計算の効率や仮説の精度は, 使用する探索戦略に大きく依存する.

3. 未知パタンと例の生成

本節では, 実験で用いる未知パタンの集合, 正例および負例の集合, 例集合に加えるノイズについて説明する.

3.1 未知パタン集合

推論の目標となるパタンの集合を**未知パタン集合**といい, P_* で表す. P_* の生成は, パタン長 l とひとつのパタンが含む変数の個数 v, パタンの個数 h を入力パラメタとして, つぎのようにおこなう:長さがちょうどl の文字列 $p \in (\Sigma \cup \{*\})^+$ で, 変数 $*$をちょうど v 個含むものを等確率に生成する; P_* は, このようなパタンをちょうど h 個含むものとする.

3.2 正負例集合

正例の多重集合 Pos は, 同じ長さの例に対しては等確率で生成する. 具体的には, 例の長さを与えるために確率分布 $Pb(\lambda, d) = (e^{-\lambda}\lambda^d)/d!$ を用いる(**Fig.4**). これは生成する例の長さが dとなる確率であり, λ は長さの期待値である. ただし, 最大の長さを 50 に制限する. また $\lambda = \infty^{50}$ に対しては, $Pb(\lambda, d) = 1/50$ と定義する.

この $Pb(\lambda, d)$ を用い, $N = \sharp Pos$ と $\lambda \geq 0$ を入力パラメタとして, つぎのように Pos を生成する:

(0) $Pos = \phi$ で初期化し, (1) から (3) を $N = \sharp Pos$ になるまで繰り返す;
(1) 未知パタン $p_* \in P_*$ を等確率に選ぶ;
(2) 例の長さ d を $Pb(\lambda, d)$ にしたがって決定する;
(3) 言語 $L(p_*)$ に含まれる長さ d の例を等確率に生成し, それを Pos に加える.

負例の多重集合 Neg の生成においても, 例の長さの決定に $Pb(\lambda, d)$ を用い, 長さ d の負例 $s \in \Sigma^+ - L(P_*)$ を等確率で生成する.

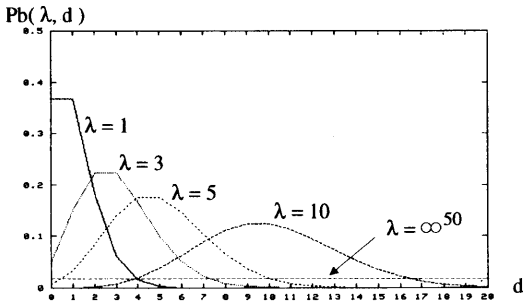

Fig.4 Probability distribution $Pb(\lambda, l)$

3.3 分類ノイズと観測ノイズ

例集合 S と実数 $0 \leq \eta \leq 1$ に対して, パラメタ η の**分類ノイズ**とは, S の正例 s が確率 η で負例となっている

ようなノイズである. 実験では, $n = \sharp S$ とパラメタ $r \leq n$ に対して, S から r 個の正例をのぞき, r 個の負例を加える.

実数 $0 \leq \delta \leq 1$ に対して, δ の**観測ノイズ**とは, 各例 $s = a_1 a_2 \ldots a_d \in S$ とすべての $1 \leq i \leq d$ に対して, 確率 δ で a_i を $\Sigma - \{a_i\}$ 中の文字に等確率で置き換えるノイズである.

4. 実験方法

本節では, MMG に対する評価実験の方法について説明する. 以下で**訓練例**とは, 学習アルゴリズムに与える正例のことである.

4.1 実験システム

未知パタン集合を P_*, 正例集合と負例集合を, それぞれ, Pos, Neg で表す. MMG の出す仮説を P で表す. 極大被覆戦略, 極小被覆戦略, 確率的戦略を, それぞれ, max, min, $rand$ で表す.

正例と負例の数を $\sharp Pos = \sharp Neg = 2000$ とする. 生成した P_* に対して, 訓練例の数 n を変化させて 20 回づつ学習する. これを, 5 回くりかえす. 仮説の精度とノイズの排除率は,それら 100 回の学習の平均値をとる.

実験システムには, 入力パラメタとして, (1) 定数文字列 Σ, パタン数$\sharp P_*$, 未知パタン $p \in P_*$ のパタン長 $l = |p|$ と変数の個数 $v = \sharp var(p)$, (2) 例の個数 $N = \sharp Pos = Neg$ と確率 $Pb(\lambda, d)$ の引数 λ, (3) 分類ノイズの個数 r と観測ノイズの置換確率 δ, (4) 訓練例の数 n, (5) 仮説 P のパタン数の上限 $k \geq \sharp P$, 仮説パタン $p \in P$ の変数の個数の上限 $m \geq \sharp var(p)$, 戦略 $\mathcal{S} \in \{max, min, rand\}$,ノイズの排除を許す例外の最大数 e を与える.

システムの構成は, (1) 3.1節で定義した確率分布にしたがって, 未知パタン集合 P_* を生成する, (2) 3.2節で定義した確率分布にしたがって, P_* から Pos と Neg を生成する, (3) ここで必要なら Pos にノイズを加える, (4) 訓練例の集合S の例を Pos から無作為に n 個選ぶ, (5) S を MMG に一括して与え, 仮説 $P \in \mathcal{RP}^k_{\varepsilon,m}$ を計算する, (6) Pos とNeg を用いて Pの精度を計算し, 必要ならP_* とS を用いてノイズの排除率を調べる, という手順からなる.

特に指定しない限り, $\sharp Pos = \sharp Neg = 2000$, $\sharp \Sigma = 5$, $\sharp P_* = k = 3$, $l = 10$, $v = m = 3$, $\lambda = 10$, $e = 0$, $n = 10 \sim 200$ とし, ノイズは含まないとする. ノイズを含む場合は, $e = 2, 4$, $r = 20$, $\delta = 0.2, 0.15$ とする.

4.2 仮説の評価方法

Pos と Neg を評価用の正負例集合とする. 仮説 P に対して, その精度を $ACC(P, Pos, Neg) = \sqrt{pos \cdot neg}$ と定義する. ただし, 正負例の精度 pos とneg をつぎのように定義する.

$$pos = \frac{\sharp(Pos \cap L(P))}{\sharp Pos}, \qquad neg = \frac{\sharp(Neg - L(P))}{\sharp Neg}$$

P_* と S を, それぞれ, 未知パタン集合と訓練例の集合とする. 仮説 P に対して, ノイズの排除率を, $EXC(P, P_*, S) = exc/noi$ と定義する. ここに, noi は S にノイズとして含まれる負例の個数であり, exc は MMG が例外として排除した負例の個数である.

5. 結果および考察

本節では, 学習アルゴリズムの評価実験の結果を示し, それを考察する.

5.1 実験 1: 未知パタン数による精度の変化

未知パタン集合のパタン数を変えて, 仮説の精度の変化を調べた. $k = \sharp P_*$, $\mathcal{S} = max$ とし, $\sharp P_* = 1 \sim 8$ で, $n = 10 \sim 200$ と変化させて実験した. 実験結果を**Fig.5** に示す.

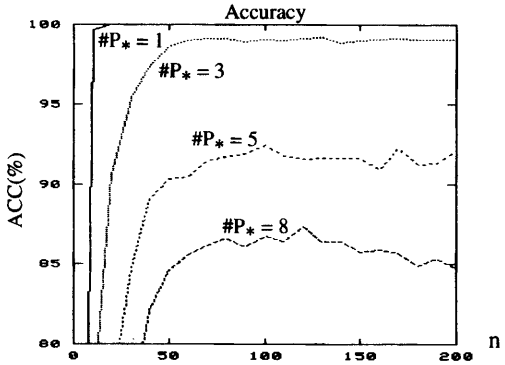

Fig.5　Accuracy for $\sharp P = 1, 3, 5, 8$

$\sharp P_* = k$ が小さいほど精度がよかった. さらに, $n = 50$ のとき, $\sharp P_* = 1, 3, 5, 8$ のそれぞれの精度は, 約 100(%), 98(%), 90(%), 85(%) であった. 以上より, MMG は,有限個の例に対しても高い精度の仮説を生成することがわかる.

5.2 実験 2: 長さの確率分布による精度の変化

例の長さの確率分布を変えて, 仮説の精度の変化を調べた. $\mathcal{S} = max$ とし, $\lambda = 5, 10, 20, \infty^{50}$ で, $n = 10 \sim 100$ と変化させて実験した. 実験結果を **Fig.6** に示す.

正例集合に短い例が多く含まれるほど, 仮説の精度がよかった. 逆に, 平均長が長い場合の精度を上げるには, n の値を大きくする必要があった. したがって, 短い例を多

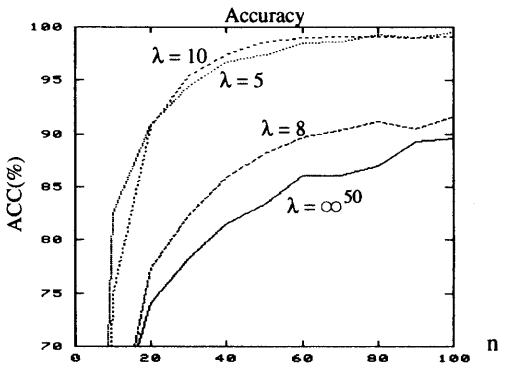

Fig.6 Accuracy for $\lambda = 5, 10, 20, \infty^{50}$

く含む正例からの学習は易しいことがわかる.

5.3 実験 3: 探索戦略の種類による精度の変化

探索戦略の種類による精度を調べた. $\mathcal{S} = max$, min, $rand$ で, $n = 10 \sim 200$ と変化させて実験した. 実験結果を **Fig.7** に示す.

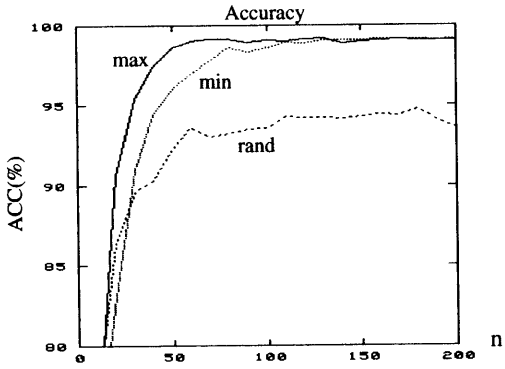

Fig.7 Accuracy for $\mathcal{S} = max, min, rand$

極大および極小被覆戦略は, 確率的戦略よりも精度がよかった. どの戦略も例数n が大きいほど精度がよく, グラフは高原状の形になった.

5.4 実験 4: 分類ノイズを含む場合の精度の変化

分類ノイズを含む場合に, 戦略の種類を変えて仮説の精度を調べた. 2000 個の正例からなる集合 Pos に 1(%) 程度の分類ノイズを加えた. $e = 2$, $r = 20$とし, $\mathcal{S} = max$, min, $rand$ で, $n = 10 \sim 300$ と変化させて実験した. 実験結果を **Fig.8** に示す.

例数が $n \geq 70$ のほとんどの場合, 極小被覆戦略, 極大被覆戦略, 確率的戦略の順に精度がよかった. ノイズの排除率を調べたところ, 極小被覆戦略は,他の戦略の 3 倍以上の割合でノイズを排除していた. したがって, この戦略の精度がよいのは, 例外としてノイズを効率良く排除して

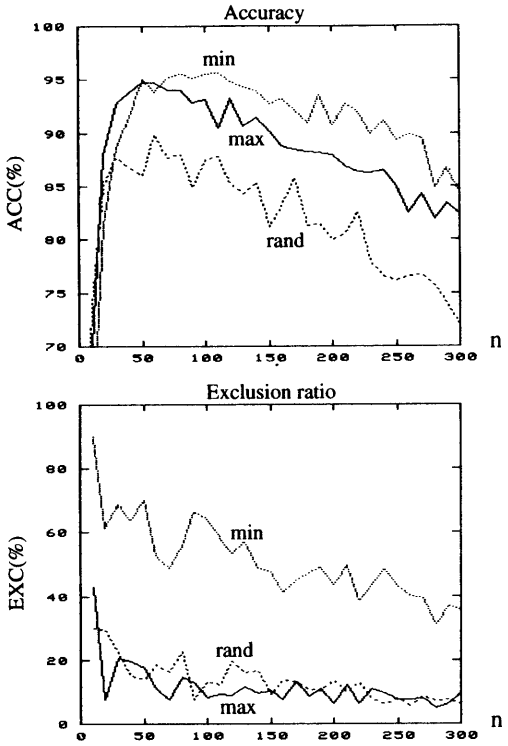

Fig.8 Accuracy and Exclusion ratio (Classification noise)

いるためと考えられる.

ノイズを含まない場合とは対象的に, 訓練例の数 n が, 極大被覆戦略と確率的戦略で 60, 極小被覆戦略で110 を超えると, どの戦略の精度も降下した.

5.5 実験 5: 観測ノイズを含む場合の精度の変化

観測ノイズを含む場合に, 戦略の種類を変えて仮説の精度を調べた. $e = 4$, $\delta = 0.2$ とし, $\mathcal{S} = max$, min, $rand$ で, $n = 10 \sim 300$ と変化させて実験した. 実験結果を **Fig.9**に示す.

例数が $n \geq 100$ の場合, 確率的戦略の精度が最も悪く, 極小被覆戦略と極大被覆戦略にはあまり差がなかった. 排除率を調べると, 極小被覆戦略が約20(%) でノイズを排除しているのに対し, 極大被覆戦略と確率的戦略は, ほとんど排除していなかった. したがって, 実験 4 の結果との比較から, 極大被覆戦略は, 分類ノイズよりも観測ノイズに強い戦略であることがわかる.

訓練例の数 n が, 大きくなると, 実験 4 と同様に, どの戦略の精度も降下した.

5.6 実験 6: ノイズの種類による精度の違い

ノイズの種類による仮説の精度を調べた. $\mathcal{S} = min$, ノイズの種類を分類ノイズと観測ノイズとし, $n = 10 \sim 300$ と変化させて実験した. 例外とノイズのパラメタは,

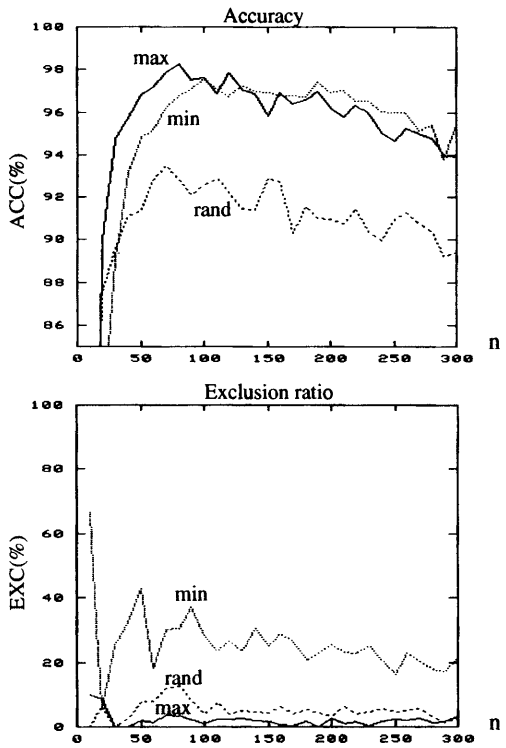

Fig.9 Accuracy and Exclusion ratio (Observation noise)

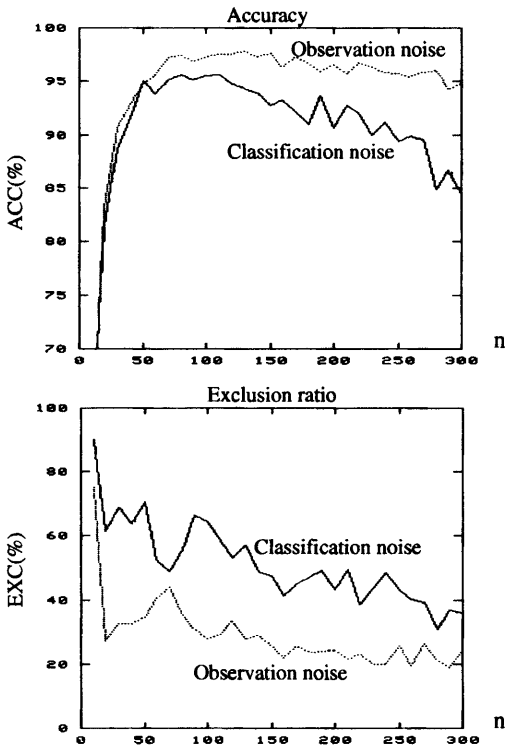

Fig.10 Classification noise and Observation noise

$e = 2, r = 20, \delta = 0.15$ とした. ただし, 観測ノイズの実験では, 正例に加えられた負例の数を分類ノイズの場合と等しくするために, 負例をちょうど 20 個含むような正例集合を選んだ. 実験結果を**Fig.10**に示す.

観測ノイズより分類ノイズの方が, ノイズが高い割合で排除されている. それにもかかわらず, 観測ノイズの方が分類ノイズよりも精度がよかった.

6. おわりに

本稿では, 確率的に生成した例からの学習実験により, MMG アルゴリズムを評価した. その結果, つぎのことがわかった.

- 長い例が多いほど, 多くの例を必要とする(実験 2).
- 探索戦略の有効性は, ノイズの有無によって異なる(実験 3 , 4, 5).
- 確率的戦略は, 人工生成データに弱い(実験 3, 4, 5).
- 分類ノイズよりも観測ノイズの方が学習が易しい(実験 6).
- ノイズ有りの場合の精度のグラフは, ゲノム実験で得られる結果[2),5)]のようにピークのある右下がりの形状をもつ(実験 1, 3, 4, 5).

本稿で扱った未知パタン集合は, 長さが等しく, 変数の個数も等しいパタンの集合であった. しかし, ゲノム情報などの現実データは, このような単純なパタン和で表すことができるとは限らない. そこで, 現実データに対する MMG の有効性をさらに解析するために, より複雑なパタン和のクラスに対してのふるまいを明らかにすることが, 今後の重要な課題である.

参考文献

1) D.Angluin. Inductive inference of formal language from positive data. *Information and Control* **45**, pp.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*, pp.39-48, 1994.
3) H.Arimura, T.Shinohara, S.Otuki. Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data. In *Proc. the 11th STACS*, LNCS 775, Springer-Verlag, pp.646-660, 1994.
4) R.Fujino. Learning unions of extended regular pattern languages from positive data and its application to discovering motifs in proteins. Master thesis, *Kyushu Inst. of Tech.*, 1994.
5) 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 38-5, pp.33-40, 1995.
6) T.Shinohara. Polynomial time inference of extended regular pattern languages. In *RIMS Symposia on Software Science and Engineering*, LNCS 147, pp.115-127, Springer-Verlag, 1982.