数理解析研究所講究録
1148 巻 2000 年 76-81

$O(\log n)$ 長の単調単項式の負例学習について

名古屋大学人間情報学研究科　築地立家 (Tatsuie Tsukiji)

名古屋大学人間情報学研究科　徳谷崇 (Takashi Tokutani)

概要

最近，遺伝子解析等への応用上の興味から，莫大な変数量を含むデータから希少変数を発見してそれらのデータを説明する問題が研究されている．本稿では，n-bit の負例集合が与えられたとき，それらと無矛盾な $O(\log n)$ 長の単調単項式を（そのような仮説が存在しない場合はその事実を）発見する問題の計算複雑さと学習アルゴリズムについて考察する．この問題は β_2 と呼ばれる計算量クラスに属する．本稿では，さらに，もしもこの問題が多項式時間で解ければ（NP 完全問題である）SAT が，$2^{O(\sqrt{n})}$ 時間でとけることを示す．また，ある（未知の）$O(\log n)$ 長の単調単項式の負例集合から一様ランダムに発生する例題を $O(\log n)$ 長の単調単項式で学習するアルゴリズムを設計する．

1　はじめに

本稿では，多くの負例と 1 つの正例（すべてが 1 である例）からなるサンプルを解析することにより，大量の変数の中から希少な重要変数を特定する問題を考察する．とくに，全変数 n に対して 関係変数の個数が $r=O(\log n)$ の場合を取り上げる．この問題の応用としては，例えばゲノム解析における遺伝子情報探索や Web 検索における鍵の絞込みなどが考えられる．そこでは，データが含む見かけ上の変数の個数は莫大であるのに対して，未知関数自体は非常に限られた個数の変数で記述できることが多い．本稿では，与えられたサンプルから関係変数の集合を正確に特定する問題，言いかえると冗長な変数を完全に除去する問題を考えたい．ただし，簡単のため，仮説は単調単項式，すなわち高々 $r=O(\log n)$ 個の正変数の連言とし，また，サンプルと完全無矛盾かつ最小の仮説を発見することを目指す．詳しくは，自然数 $r=O(\log n)$ が与えられたとき，もしサイズ r 以下の無矛盾仮説が存在すればその中のひとつを発見し，存在しなければその事実をつきとめなければならない．これを **CONJ**$(O(\log n), n)$-仮説での負例学習と呼ぶ．

本稿では，まず任意に与えられた負例集合の学習困難性について考察する．r に何らの制限を設けなければ（例えば $r=n/2$ であれば）この問題は SET COVER の解発見問題と同等となり，よって NP 困難である.（2 章参照）．この事実に基づき，Dhagat-Hellerstein[2] は，仮説が $O(\log n)$ 倍冗長であることを許せば Greedy Algorithm がこの問題を多項式時間で解くことと，その詳しい性能の分析を行っている．本稿では，特に $r=O(\log n)$ である場合の計算困難性を考察する．この問題は β_2 と呼ばれる計算量クラスに属する．なぜなら，$O(\log^2 n)$ ビットの推測により関係変数の集合 $X_{i_1},\dots,X_{i_r}$ を特定し，その後多項式時間の計算で，$\mathrm{AND}(X_{i_1},\dots,X_{i_r})$ が与えられた例題をすべて負例に分類することを確認すればよいからである．

本稿では，

> もしも **CONJ**$(O(\log n), n)$-仮説での負例学習が多項式時間で可能であれば,（代表的な NP 完全問題である）SAT が，$2^{O(\sqrt{n})}$ 時間以内で決定的に解ける

ことを示す．

また，本稿では，負例集合が任意に固定された未

知関数 $f \in \mathbf{CONJ}(r,n)$ の負例全体のなかから一様にサンプルされるとき f を学習する問題も考察する．まず，

> f を唯一に特定するために必要十分なサンプルサイズは，$\Omega(2^r \log n)$ 以上 $O(r2^r \log n)$ 以下である

ことを明らかにする．また，

> $O(4^r \log n)$ 個のサンプルから $O(r4^r n \log n)$ 時間で f を発見できる．

このアルゴリズムは各変数のもつ例題との相関値を計算し，相関値の高い変数の集合を出力する．

2 準備と背景

$\mathbf{CONJ}(r,n)$ を n 個の変数のうち r 個以下の変数で構成される単調単項式のクラスとする．よって，$X_1, \ldots, X_n$ を全変数とするとき

定義 1 $\mathbf{CONJ}(r,n) =$

$$\{\mathrm{AND}(X_{i_1}, \ldots, X_{i_k}) \ : \ i_1 < i_2 < \cdots < i_k, k \leq r\}$$

である．

以降は便宜上 n は r の倍数であるとして，B-CONJ(r,n) は $\{X_1, \ldots, X_n\}$ を r 分割した各ブロック $B_j = \left\{X_{(j-1)n/r+1}, \ldots, X_{jn/r}\right\}$ からちょうど一個づつ変数を選んでできる単調単項式のクラス，すなわち

定義 2 B-CONJ$(r,n) =$

$$\{\mathrm{AND}(X_{i_1}, \ldots, X_{i_k}) \ : \ X_{i_j} \in B_j\}$$

とする．

また単調単項式 f に対して rel(f) を f を構成している変数からなる集合とする．

ブール関数 f の入力 I（例題とも呼ぶ）は $f(I) = 0$ のとき負例と呼ばれ，また $f(I) = 1$ のとき正例と呼ばれる．$\mathcal{C}$ および $\mathcal{H}$ をブール関数のクラスとする．このとき，次のような 2 つの問題を考える．

問題 1 $\mathcal{H}$-仮説での負例学習問題とは，任意に与えられた例題集合に対し，それらをすべて負例に分類するような $h \in \mathcal{H}$ を発見するか，または，そのような h が存在しない場合には，そのことを判定する問題である．

上の問題に対し，ある $f \in \mathcal{C}$ の存在を仮定した問題も考える．すなわち，

問題 2 $\mathcal{C}$-関数の $\mathcal{H}$-仮説での負例学習問題とは，任意に固定された $f \in \mathcal{C}$ の負例集合をすべて正しく（すなわち負例に）分類するような $h \in \mathcal{H}$ を発見する問題である．

$\mathbf{CONJ}(r,n)$-仮説での負例学習は r-点での集合被覆問題の双対問題である．よって，特に

定理 1 (Dhagat[2]) $\mathbf{CONJ}(r,n)$-仮説での負例学習は NP-完全である．

また，集合被覆問題の近似解を求める Greedy Algorithm を使うと，

定理 2 (Dhagat[2]) $\mathbf{CONJ}(r,n)$-関数の $\mathbf{CONJ}(O(r \log n), n)$-仮説での負例学習は多項式時間計算可能である．

Kintala-Fischer [4] は NP-機械の非決定性ステップの回数を polylog(n) に制限することにより，β-階層とよばれる計算量のクラスの階層を用意した．

定義 3 β_k とは，ある多項式時間計算可能な 2 項述語 R を用いて
$\left\{I \ : \ R(I,w) = 1 \textit{ for some } w \textit{ with } |w| \leq \log^k |I|\right\}$
と表現できる言語のクラスである．

特に，$\mathbf{CONJ}(O(\log n), n) \in \beta_2$ である．Szelepcsényi [3] は many-one 還元に関して β_k 完全な集合の存在を証明した．ただし Szelepcsényi の完全集合はそれを解くための Greedy Algorithm と対の形で定義される．例えば与えられた 3-CNF 式 $F(X_1, \ldots, X_n)$ と $X_1, \ldots, X_{O(\log^k n)}$ への真偽値割り当ての組 (F, A) から充足割当てを発見する Greedy Algorighm とは，次の 2 つのステップを定常状態に至るまで繰り返すことである．

step-1 同じ節の中の1つ以外のすべてのリテラルの値が0であり，残りのリテラルの値が決定していないのであれば，そのリテラルに1を与える．

step-2 step-1の割当てを全てのcaluseに反映させる．

このとき，

定義 4 A を $X_1, \dots, X_{O(\log^k n)}$ への真偽値割当てとすると，

SAT$_k$=$\{F \in$ CNF : $(\exists A)$ Greedy Algorithm は (F, A) から充足割り当てを発見する $\}$.

定理 3 (Szelepcsényi[3]) 任意のkに対し，SAT$_k$ は β_k-完全である．

3 学習困難性

n 変数で s 個の節から成る3-SAT式の中で充足可能な式の集合を 3-SAT(n, s) と表す．$r = O(\log n)$, $p = n^{O(1)}$ をそれぞれ任意に固定し，便宜上 $n = r2^k$, $k > 0$ は整数であるとする．

定理 4 3-SAT$(kr, p(n))$ は B-CONJ(r, n) の負例学習問題に $n^{O(1)}$-時間で many-one 還元される．

(証明) 便宜上，3-SAT$(kr, p(n))$ 式の表記に用いる kr 個の変数を $x_i^{(j)}$, $1 \le j \le r$, $0 \le i < k$, とかく．また，B-CONJ(r, n) 式の表記に用いる n 個の変数を $X_a^{(j)}, 1 \le j \le r$, $0 \le a < 2^k$ とかく．また a の2進表示を $a_0a_1 \dots a_{k-1} \in \{0,1\}^k$, $a = a_0 + 2a_1 + 4a_2 + \dots + 2^{k-1}a_k$, とする．さて，与えられた 3-SAT$(kr, s)$ 式

$$F = C_1 \wedge C_2 \wedge \dots \wedge C_s$$

の各節 C_m を用いて $X_1, \dots, X_n$ への真偽値割り当て $I_m \in \{0,1\}^n$ を次のように構成する．

$$X_a^{(j)}(I_m) = \begin{cases} 0 & \text{if either} \\ & \quad a_i = 1 \text{ and } x_i^{(j)} \in C_m \\ & \text{or} \\ & \quad a_i = 0 \text{ and } \neg x_i^{(j)} \in C_m, \\ 1 & \text{otherwise.} \end{cases}$$

例 $n = 16, r = 2, k = 3$ のとき
節 $\left\{x_0^{(1)}, \neg x_1^{(1)}, x_2^{(2)}\right\}$ は，
$(0,0,1,0,0,0,1,0,1,1,1,1,0,0,0,0)$ である．

補題 1 任意の 3-CNF$(kr, p(n))$ 式 F とそれから構成した真偽値割り当て $I_1, \dots, I_s$ について，F が充足可能であるときかつそのときに限って $I_1, \dots, I_s$ で偽となるような B-CONJ(r, n) 式が存在する．

(証明) まず，F の充足割当て $A = (A^{(1)}, \dots, A^{(r)}) \in \{0, 1, \dots 2^k - 1\}^r \cong \{0,1\}^{kr}$ が与えられたとき，$f = \mathrm{AND}(X_{A^{(1)}}^{(1)}, \dots, X_{A^{(r)}}^{(r)})$ とおけば，$f(I_1) = \dots = f(I_s) = 0$ となることをしめす．任意の節 C_m に対して，$C_m(A) = 1$ なのであるリテラル（変数またはその否定）$y \in C_m$ に対して $y(A) = 1$ となり，さらに y が第 j ブロックに属するとき $y(A^{(j)}) = 1$ となる．よって，I_m の構成から $X_{A^{(j)}}^{(j)}(I_m) = 0$ となり，$f(I_m) = 0$ をえる．

逆に，$f(I_1) = \dots = f(I_s) = 0$ なる $f \in$ B-CONJ(r, n) が与えられたとき，$f = \mathrm{AND}(X_{A^{(1)}}^{(1)}, \dots, X_{A^{(r)}}^{(r)})$, $A \in \{0,1\}^{kr}$, とかける．ゆえに，任意の負例 I_m に対してある変数 $X_{A^{(j)}}^{(j)}$ が存在して $X_{A^{(j)}}^{(j)}(I_m) = 0$ となる．したがって，I_m の構成法により，ある C_m のリテラル $y \in \{x_i^{(j)} \neg x_i^{(j)}; 0 \le i < k\}$ が存在して，$y(A^{(j)}) = 1$ となる．よって，任意の節 C_m に対して $C_m(A) = 1$ となるので，$F(A) = 1$ を得る．□

系 1 もしも B-CONJ$(O(\log n), n)$-仮説での負例学習が $n^{O(1)}$ 時間で実行可能であれば，3-SAT$(n, 2^{O(\sqrt{n})})$ が $2^{O(\sqrt{n})}$ 時間で認識可能となる．

(証明) $k = r = \sqrt{n}$, $n = r2^k$. 仮定より B-CONJ(r, n) は $2^{O(\sqrt{n})}$ 時間で負例学習可能である．よって定理4より 3-SAT$(kr, p(n))$ = 3-SAT$(n, 2^{O(\sqrt{n})})$ は $2^{O(\sqrt{n})}$ 時間で認識可能となる．□

4 一様分布における負例学習

任意に固定された未知関数 $f \in$ **CONJ**(r, n) の負例全体 $\{I \in \{0,1\}^n; f(I) = 0\}$ から一様ランダム

かつ独立に例題$I_1,\ldots,I_m$を選ぶとき，fを唯一に特定する問題を考える．そのためには，

> [**uniqueness**] 任意の$h\in\mathbf{CONJ}(r,n)$について$h(I_1)=\cdots=h(I_m)=0$ならば$h=f$

でなければならない．そこでまず，**uniqueness**が高い確率で成立するためのmの下界と上界を与える．

補題 2 実数$\delta>0$と整数$m\geq 2^r(r\log n+\log\frac{1}{\delta})$と関数$f\in\mathbf{CONJ}(r,n)$を任意に固定する．このとき，**uniqueness**が成立する確立は$1-\delta$以上である．

（証明） $f,h\in\mathbf{CONJ}(r,n)$を各々任意に固定する．$|\mathrm{rel}(h)\cap\mathrm{rel}(f)|=i$, $0\leq i\leq r$, とする．また，fの負例Aを一様ランダムに取り出す．このとき，

$$\begin{aligned}\mathrm{Prob}_A\{h(A)=1\} &= \frac{2^{n-2r+i}(2^{r-i}-1)}{2^{n-r}(2^r-1)}\\ &= \frac{2^{r-i}-1}{2^{r-i}(2^r-1)}\end{aligned}$$

である．したがって負例を独立にm個とるときそのどれとも$h(A)=1$とならない確率は

$$\left(1-\frac{(2^{r-i}-1)}{2^{r-i}(2^r-1)}\right)^m$$

となる．従って，f以外のすべての仮説についてこれが成立しない確率は

$$\begin{aligned}&\sum_{i=0}^{r-1}\binom{r}{i}\cdot\binom{n-r}{r-i}\cdot\left(1-\frac{(2^{r-i}-1)}{2^{r-i}(2^r-1)}\right)^m\\ &< \binom{n}{r}\cdot\left(1-\frac{1}{2^{r-1}}\right)^m\end{aligned}$$

で押さえられる．ここで

$$m\geq 2^r\left(r\log n+\log\frac{1}{\delta}\right)$$

であれば，最終右辺はさらにδで押さえられる． □

$X=\mathrm{Binary}(S,p)$をサイズS, 確率パラメータpの2項分布とする．よって$\mathrm{Prob}X=k=\binom{S}{k}p^k(1-p)^{S-k}$である．

補題 3 任意の$\varepsilon>0$について

$$\mathrm{Prob}\{X\geq(p+\varepsilon)S\}<e^{-2\varepsilon^2S}$$

と

$$\mathrm{Prob}\{X\geq(1+\varepsilon)pS\}<e^{-\varepsilon^2pS/3}$$

が成立する．

補題 4 X,Yを実確率変数とする．任意の実数rについて

$$\mathrm{Prob}\{X\geq r\}\leq\mathrm{Prob}\{Y\geq r\}$$

が成立すれば，ある$\mathrm{dom}(X)\times\mathrm{dom}(Y)$上の確率分布のもとで$X\leq Y$が確率1で成立する．

補題 5 $X=\mathrm{Binary}(S,p)$, $Y=\mathrm{Binary}(T,q)$, を2項分布とする．$S\leq T$, $p\leq q$ならば，ある確率分布のもので$X\leq Y$が確率1で成立する．

（証明） 任意の$r\geq 0$について

$$\begin{aligned}\mathrm{Prob}\{X\geq r\} &= \sum_{i\geq r}\binom{S}{i}\cdot p^i(1-p)^{S-i}\\ &\leq \sum_{i\geq r}\binom{T}{i}\cdot q^i(1-q)^{T-i}\end{aligned}$$

なので 補題4より． □

補題 6 実数$\varepsilon,\delta>0$と整数$m\leq(2^r-1)\bigl(\ln 2(n/r)\varepsilon^2-\ln\ln(4/\delta)\bigr)/\ln(2(1-2\varepsilon))$と関数$f\in$ B-CONJ を任意に固定する．**uniqueness**が成立する確立はδ以下である．

（証明） $f=\mathrm{AND}(X_{i_1},\ldots,X_{i_r})$, $X_{i_j}\in B_j$とする．さて，X_{i_1}がfの変数であることを確認するための負例Aは

$$A(X_{i_2})=\ldots=A(X_{i_r})=1 \qquad (1)$$

を満たさねばならない．そこで，負例Aを一様ランダムにT個とってきたときに，(1)をみたすものの個数をXであらわすと，Xの期待値は

$$\mathrm{E}[X]=T/(2^r-1)$$

であり，さらに各負例は独立に選択されるので，補題 3 より任意の $0<\varepsilon<1/2$ について

$$\mathrm{Prob}\{X \geq T(1+\varepsilon)/(2^r-1)\} < e^{-T\varepsilon^2/(3(2^r-1))}$$

が成立する．例えば，与えられた $j_0 \geq (3/\varepsilon^2)\ln(2/\delta)$ について $T = j_0\cdot(2^r-1)$ とおくと

$$\mathrm{Prob}\{X \geq j_0(1+\varepsilon)\} < e^{-j_0\varepsilon^2/3} < \delta/2 \qquad (2)$$

である．

次に，(1) をみたす負例を一様ランダムに分布させる．このとき，$S_0 = \{x_1,\ldots,x_{n/r}\}$, $S_j = \{x_i \in S_{j-1} : A(x_i)=0\}$ for $j>0$ とすると，S_j は j がひとつ増えるにつれて半分にへると思われる．実際，$S_{j-1}=S$ の条件のもとで期待値は

$$\mathrm{E}[|S_j| \mid S_{j-1}=S] = \frac{|S_{j-1}|-1}{2}$$

である．また，任意の $0<\varepsilon<1/2$ について，補題 3 より

$$\mathrm{Prob}\{|S_j| < (1/2-\varepsilon)|S| \mid S_{j-1}=S\} \leq e^{-2\varepsilon^2|S|}$$

が成り立つので，補題 5 より

$$\mathrm{Prob}\big\{|S_j| < (1/2-\varepsilon)^j\cdot n/r\big\} < \sum_{i=0}^{j-1} e^{-2\varepsilon^2(1/2-\varepsilon)^i n/r}$$

となる．そこで

$$j_0 = \frac{\ln(2\varepsilon^2 n/r) - \ln\ln(4/\delta)}{\ln(2/(1-2\varepsilon))}$$

と設定すると

$$\mathrm{Prob}\big\{|S_{j_0}| < (1/2-\varepsilon)^{j_0} n/r\big\} < \delta/2$$

で，さらに

$$(1/2-\varepsilon)^{j_0}\cdot n/r = \frac{\ln(4/\delta)}{2\varepsilon^2} > 1$$

より

$$\mathrm{Prob}\{|S_{j_0}|>1\} < \delta/2 \qquad (3)$$

となる．

最後に，与えられた $0<\varepsilon<1/2$ について，$T = (2^r-1)\big(\ln 2(n/r)\varepsilon^2 - \ln\ln(4/\delta)\big)/\ln(2(1-2\varepsilon))$ とおく． すると $T < j_0(2^r-1)$ なので，(2) と (3) より，確率 $1-\delta/2-\delta/2$ 以上で $X \in B_1$, $X \neq X_{i_1}$ が存在して，$\mathrm{AND}(X, X_{i_2},\ldots,X_{i_r})$ はランダムに選んだ T 個の負例すべてと無矛盾となる．
□

$r=O(\log n)$ の時に補題 2 の上界と 補題 6 の下界はともに n の多項式であることを注意する．最後に，サンプルサイズ m が **uniqueness** を保証するくらい十分与えられたとき効率的に f を発見するアルゴリズムを紹介する．

LearnConj

入力: 例集合 S と実数 $\varepsilon>0$.

出力: 変数の集合 V

初期設定 $V := \emptyset$, $m := 0$, $l = 0$.

For $i = 1,\ldots,n$

$|I \in S : X_i(I) = 0| \geq (1/2)(1+\varepsilon)|S|$ ならば，X_i を V の中に入れる．

定理 5 実数 $\varepsilon \leq (1/2)(2^r/(2^r-1)-1)$，整数

$$m \geq (2/\varepsilon^2)(\log n + \log(1/\delta))$$

と関数 $f \in \mathbf{CONJ}(r,n)$ を任意に固定し，S を f の負例全体から一様ランダムかつ独立に m 個サンプルした例集合とする．このとき，**LearnConj** は確率 $1-\delta$ 以上の確率で $\mathrm{rel}(f)$ を出力する．

(証明) f の負例を m 個取るとき，任意の $X \in \mathrm{rel}(f)$ について，X に 0 を割り当てる負例の数が $(m/2)(1+\varepsilon)$ 個より多くでる確率は，補題 3 より

$$1 - \mathrm{Prob}\{S < (1+\varepsilon)m/2\} > 1 - e^{-(m/2)\varepsilon^2}$$

以上であり，また任意の $Y \in \{X_1,\ldots,X_n\} - \mathrm{rel}(f)$ について，X に 0 を割り当てる負例の数が $\frac{m}{2}(1+\varepsilon)$ 個より少なくでる確率は $1-e^{-(m/2)\varepsilon^2}$ 以上である．従って，V を **LearnConj** の出力とすると，全ての変数 X について

$$X \in \mathrm{rel}(f) \longleftrightarrow X \in V$$

が成り立つ確率は

$$1 - ne^{-(m/2)\varepsilon^2} > 1 - \delta$$

となる． □

5 おわりに

$\mathbf{CONJ}(r, n)$-仮説での負例学習が β_2 に属することはすでに述べたが，β_2-完全かどうかはわからない．Buss-Goldsmith [1] らは $\mathrm{SAT}(\log^2 n, \mathrm{polylog}(n))$ が β_2-完全ではないことを指摘している．彼らの証明を拡張すれば，$\mathrm{SAT}(\log^2 n, \mathrm{poly}(n))$ すらも β_2-完全でないことがわかるので，$\mathbf{CONJ}(O(\log n), n)$ も β_2-完全でないと予想される．

謝辞

本研究に当たり，多大な助言をくださった 渡辺教授，Carlos Doming 博士， および David Guijarro 博士に深く感謝します．

参考文献

[1] J. Buss and J. Goldsmith. Nondeterminism within P. *Proceedings of the 8th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science*, 480, 1991.

[2] A. Dhagat and L. Hellerstein. PAC Learning with Irrelevant Attributes. *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, 64–74, 1994.

[3] R.Szelepcsényi. β_k-Complete Problems and Greediness . *manuscript*, 1993.

[4] C.Kintala and P.Fisher. Refining nondeterminism in relativized polynomial-time bounded computatins. *SIAM Journal on Computing*, 9(1):46-53,1980.