(15)「最近の計算機支援型推測の基礎理論とその応用」に関する研究報告

Quantam Walks on the Contingency tables.

九州芸工大　坂田年男
岡山理科大学　澤江隆一

1. はじめに

本講演では、はじめに量子計算の基礎事項を簡単に振り返り、最近話題になっている量子ランダムウオークの基礎概念を概観した。　さらに、周辺和一定の分割表集合の上のランダムウーオークは統計学では重要なＭＣＭＣの応用分野であるが、　その量子版を提案し,極限で一様分布に収束することを量子アルゴリズムのエミュレーションで示した。

2.　直線状のランダムウオークについての概観

＜古典的酔歩＞

原点を出発する確率1/2で右または左へ進むランダムウオークは｛コインを投げる｝＋｛格子点を動く｝という２つの要素からなる.そこでは、コイン投げの結果の空間＝｛０，１｝と点の状態空間＝｛直線上の格子点全体｝が使われていると考えられる。これを量子酔歩の場合にも踏襲する。

＜直線上の量子ランダムウオーク＞

コイン投げを表す２次元ヒルベルト空間　Ｈｃ＝｛｜＋＞、｜－＞｝と格子点の集合を表すヒルベル空間　H_{s}={ |j>; j=1,2,,, -1,-2,,, }を考え、全体の状態空間は二つの空間のテンソル積とする。　量子酔歩の状態発展は量子力学の要請からユニタリー変換で記述されなければならず、ここで考えるユニタリー変換は、まずコイン投げ空間H_cの上のユニタリー変換としてアダマール変換を考える。　次に、酔歩の格子点上の推移を与えるユニタリー変換を

$$H = H_c \otimes H_s$$

$$F(|j> \otimes |+>) = |j+1> \otimes |+>,$$

$$F(|j> \otimes |->) = |j-1> \otimes |-1>$$

とし、全体の状態空間へ作用するユニタリー変換を

$$U = F \otimes (H \otimes I)$$

とする。Ｕは確かにユニタリー変換である。

＜量子酔歩の確率分布の特徴＞

Ｕをｎ回かけることは、ｎ回コインを投げることに対応する。量子酔歩は可逆であるから,古典的酔歩とは同じ状態遷移をしない。１）非対称な分布であること。２）発散距離がステップ数に比例するなどの古典論にない特徴がある。

3.　分割表の上のquantum　Markov chain

文献的にはグラフ上の量子マルコフ連鎖の論文がすでに出ている。
ここで、紙数の関係でその定義を振り返ることはできないので直接分割表の上の酔歩に入る。

＜周辺和一定の分割表の集合上の古典酔歩＞

簡単のため、２元表を考える（高次の場合も構想はできている）。古典的酔歩はランダムな行ペアー(I,I)とランダムな列ペアー(J,J)を選び、確率1/2で

	J	J'		J	J'
I	+1	−1	I	−1	+1
I'	−1	+1	I、	+1	−1

のどちらかを作用させる。進めない時は立ち止まり、再試行する。

＜分割表上の酔歩の量子版＞

まず、コイン投げの空間H_{C1}＝｛+、−｝、行選択空間H_{C2}＝｛(1,2),(1,3),,,(k-1,k)｝
列選択空間H_{C3}＝｛(1,2),(1,3),,,(L-1,L)｝、条件を満たす分割表全体の空間H_sを用意する。

提案するユニタリー変換としては、

1）コイン投げの空間H_{C1}＝｛+、−｝にはランダムアダマール変換

2）　行選択空間H_{C2}＝｛(1,2),(1,3),,,(k-1,k)｝

3）　列選択空間H_{C3}＝｛(1,2),(1,3),,,(L-1,L)｝
　　には量子フーリエ変換をそれぞれに施す。

4）条件を満たす分割表全体の空間H_sには

$$F(+,(I,I'),(J,J'),v)=\begin{cases}(-,(I,I'),(J,J'),v) & \textit{if } v' \textit{ is unreachable}\\ (+,(I,I'),(J,J'),v') & \textit{if } v' \textit{ is reachable}\end{cases}$$

where v' is the table obtained by v when +,(I,I'),(J,J') is operated

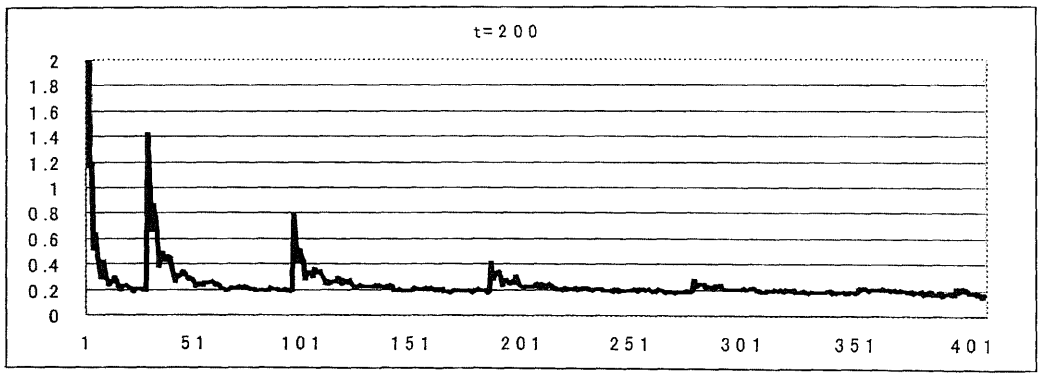
$$F(-,(I,I'),(J,J'),v)=\begin{cases}(+,(I,I'),(J,J'),v) & \textit{if } v' \textit{ is unreachable}\\ (-,(I,I'),(J,J'),v') & \textit{if } v' \textit{ is reachable}\end{cases}$$

where v' is the table obtained by v when −,(I,I'),(J,J') is operated

を作用させる。

<時間平均分布の収束状況＞

３ｘ３で周辺和がすべて６であるあるひとつの表に対して量子酔歩を施したときのステップ数に対する時間平均分布一様分布への収束状況を示す以下のデータを得た。



＜今後の課題＞

一様分布ではなくて、超幾何分布へ収束させるシステムをどう構築するかの問題が残った。
また、どのような応用が考えられるかという基本的問題も残っている。

MCMC による誤分類を含む分割表の解析

黒田 正博 (倉敷芸術科学大学ソフトウエア学科)

1 はじめに

疫学研究のようなある集団を対象とした観察研究において，調査方法等の不備により情報が正しく得られず，観測データに情報バイアスが入ってしまうことがある．このような情報バイアスによって誤分類 (misclassification) された分割表は，推定値にバイアスを与え，解析結果を歪めてしまう可能性がある．

誤分類誤差を含む分割表解析に関して，多くの研究者によって議論がなされており，これらは [1], [2], [4] に詳しい．通常，誤分類誤差を調整したセル確率の推定値は，モーメント法あるいは最尤法によって求められ，数値解法として EM 法，Fisher scoring 法が適用される．しかしながら，EM 法ではセル確率の推定値を得ることができるが，推定値の分散，信頼区間等の評価をおこなうことはできない．また，Fisher scoring 法は収束も速く推定値の分散を求めることができるが，各反復で情報行列の計算が必要であり，初期値によっては EM 法に比べ不適解を生じやすいという不安定さもある．そこで，Bayes 法によりセル確率の事後分布を求め，推定値とその分散の評価および信頼信頼区間を構成することを考える．また，Bayes 法を適用することで，誤分類に関する事前情報を結局的に推定に織り込むことが可能になる．

本稿では，Bayes 法により誤分類を含む分割表のセル確率を推定する．このとき，精密(exact) な事後分布の計算は多重積分等が必要となり困難であるので，Data Augmentation(DA) 法 [3] を用いて導出する．

2 誤分類を含む分割表の Bayes 法による推定

カテゴリカル変数 X, Y と Y を誤分類した Y' に対して，観測データ

$$n=\{n_{i+k} \mid i\in\{1,\ldots,I\}, k\in\{1,\ldots,K\}\}, \quad m=\{m_{+jk} \mid j\in\{1,\ldots,J\}, k\in\{1,\ldots,K\}\}.$$

が得られたもとでの 3 元分割表のセル確率

$$\theta_{XYY'}=\{p_{ijk} \mid i\in\{1,\ldots,I\}, j\in\{1\ldots,J\}, k\in\{1,\ldots,K\}\}$$

を Bayes 推定によって求めることを考える．ここで，記号 "+" は対応する変数の周辺和を意味する．このとき，観測データ n, m が，それぞれ $\theta_{XYY'}$ の周辺確率

$$\theta_{XY'}=\{p_{i+k} \mid i\in\{1,\ldots,I\}, k\in\{1,\ldots,K\}\}, \quad \theta_{YY'}=\{p_{+jk} \mid j\in\{1,\ldots,J\}, k\in\{1,\ldots,K\}\}.$$

をパラメータにもつ多項分布 $f(n \mid \theta_{XY'})$, $f(m \mid \theta_{YY'})$ に従うとすれば，周辺確率 $\theta_{XY'}$, $\theta_{YY'}$ の自然共役事前分布は，Dirichlet 分布であるので，確率 $\theta_{XYY'}$ の事前分布として Dirichlet 分布 $\pi(\theta_{XYY'} \mid \alpha)$ を仮定することができる．ここで，$\alpha=\{\alpha_{ijk} \mid i\in\{1,\ldots,I\}, j\in\{1,\ldots,J\}, k\in\{1,\ldots,K\}\}$ はハイパーパラメータであり，誤分類に関する事前知識を表現することができる．このとき，尤度は，2 つの多項分布の積 $f(n,m \mid \theta_{XY'},\theta_{YY'})$ であり，事後分布は事前分布と尤度の積によって求められる：

$$\pi(\theta_{XYY'} \mid n,m) \propto \pi(\theta_{XYY'} \mid \alpha) \times f(n,m \mid \theta_{XY'},\theta_{YY'}). \tag{1}$$

しかしながら，このとき事後密度の計算は非常に困難であり，データサイズがそれ程大きくないときでさえ，精密な事後分布を導出することは多大な計算が必要となる．そこで，DA 法を用いて事後分布の近似分布を求め，確率 $\theta_{XYY'}$ の Bayes 推定をおこなう．

3 DA 法による推定

DA 法は，Markov Chain Monte Carlo 法の 1 つであり，特に不完全データからの事後密度は複雑であるが，想定したモデルの完全データからの事後密度が比較的扱いやすく，標本の生成が容易な場合によく用いられる [5].

本稿で議論する多項モデルは，n, m が完全データの場合，これらを $\tilde{n}$, $\tilde{m}$ で表すと，その事後分布 $\pi(\theta_{XY^{\prime}} \mid \tilde{n}, \tilde{m})$ は，非常に簡単な密度関数によって表現できる．そこで，不完全データ n, m の擬似完全データ $\tilde{n}$, $\tilde{m}$ をDA 法により生成し，完全データの枠組みでの Bayes 推定をおこなう．DA 法は，Imputation step(I-step), Posterior step(P-step) と呼ばれる 2 つステップを交互に繰り返す．ここで考える多項モデルにおいて，DA 法は以下の手順で与えられる：

***I*-step:** 擬似完全データ

$$\tilde{n} = \{\tilde{n}_{ijk} \mid i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}, \sum_{j} \tilde{n}_{ijk} = n_{i+k}, \tilde{n}_{ijk} > 0\},$$

$$\tilde{m} = \{\tilde{m}_{ijk} \mid i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}, \sum_{i} \tilde{m}_{ijk} = m_{+jk}, \tilde{m}_{ijk} > 0\}$$

を生成するために，次の 3 つのステップを $l = 1,\ldots,L$ 回繰り返す．

1. 事後分布 $\pi(\theta_{XY^{\prime}} \mid \tilde{n}, \tilde{m})$ に従う確率 $\theta^{*}_{XY^{\prime}}$ を生成する．
2. 予測分布 $f(\tilde{n} \mid \{p^{*}_{j|i,k}\})$ に従う擬似完全データ $\tilde{n}^{(l)}$ を生成する．
3. 予測分布 $f(\tilde{m} \mid \{p^{*}_{i|j,k}\})$ に従う擬似完全データ $\tilde{m}^{(l)}$ を生成する．

   ここで，$p^{*}_{j|i,k} = p^{*}_{ijk}/p^{*}_{i+k}$, $p^{*}_{i|j,k} = p^{*}_{ijk}/p^{*}_{+jk}$ である．

***P*-step:** I-step で生成した擬似完全データ $\tilde{n}^{(l)}$, $\tilde{m}^{(l)}$ $(l = 1,\ldots,L)$ をもとに，$\theta_{XY^{\prime}}$ の近似事後分布 $\pi(\theta_{XY^{\prime}} \mid n, m)$ を更新する：

$$\pi(\theta_{XY^{\prime}} \mid m, n) = \frac{1}{L} \sum_{l=1}^{L} \pi(\theta_{XY^{\prime}} \mid \tilde{n}^{(l)}, \tilde{m}^{(l)}).$$

この I-step, P-step を，近似事後分布 $\pi(\theta_{XY^{\prime}} \mid m, n)$ が定常分布にが達するまで繰り返す．この分布が定常分布に達した時，事後分布 (1) が得られたことになる．このとき，事後平均により確率 $\theta_{XY^{\prime}}$ の推定値を得ることができる．また，事後分散，信頼区間も容易に求めることができる．さらに，推定精度を上げるためには Monte Carlo 法のアイデアにより十分に大きく L の値を設定すればよい．

参考文献

[1] Chen, T.T. (1989). A review of methods for misclassified categorical data in epidemiology, *Statistics in Medicine*, 8, 1095–1106.

[2] Kuha, J. and Skinner, C. (1997). Categorical data analysis misclassification, *Survey Measurement and Process Quality*, Ed. L. Lyberg et al., New York: Wiley.

[3] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

[4] Van den Hout, A. and Van der Heijden, P. (2000). Randomized response, statistical disclosure control and misclassification: a review, *personal communication.*

[5] 渡辺美智子，山口和範 編，(2000). EM アルゴリズムと不完全データの諸問題, 多賀出版.

Random Generation of $\mathrm{B}^m \times J$ Contingency Tables

University of Tokyo	Tomomi MATSUI	tomomi@misojiro.t.u-tokyo.ac.jp
Tokai University	Yasuko MATSUI	yasuko@ss.u-tokai.ac.jp
Science University of Tokyo	Yoko ONO	ono@ms.kagu.sut.ac.jp

1 Introduction

We propose a new Markov chain for sampling $\mathrm{B}^m \times J = \mathrm{B} \times \cdots \times \mathrm{B} \times J$ contingency tables where $\mathrm{B} = \{1,2\}$ and $J = \{1,2,\ldots,n\}$. This Markov chain is an extention of the Markov chain which is proposed by Dyer and Greenhill [3] for two rowed contingency tables. To show that our Markov chain is rapidly mixing, we use a path coupling method, which is proposed by Bubley and Dyer [1].

2 Contingency Tables

We denote the set of integers (non-negative integers, positive integers) by Z (Z_+, Z_{++}) respectively and consider a set of contingency tables indexed by $\mathrm{B}^m \times J$ where $\mathrm{B} = \{1,2\}$ and $J = \{1,2,\ldots,n\}$. Any index in J is called a *column index*. For any vector $\boldsymbol{x} \in \mathbf{R}^{\mathrm{B}^m \times J}$, both $x(\boldsymbol{i};j)$ and $x(i_1,i_2,\ldots,i_m;j)$ denote the elements of $\boldsymbol{x}$ indexed by $\boldsymbol{i} = (i_1,i_2,\ldots,i_m) \in \mathrm{B}^m$ and $j \in J$. For any column index $j \in J$, $\boldsymbol{x}(j) \in \mathbf{R}^{\mathrm{B}^m}$ denotes the subvector of $\boldsymbol{x} \in \mathbf{R}^{\mathrm{B}^m \times J}$ consists of elements defined by indices in $\mathrm{B}^m \times \{j\}$. Given a vector of indices $\boldsymbol{i} \in \mathrm{B}^m$, $\boldsymbol{i}_{\overline{l}}$ denotes the vector $(i_1,\ldots,i_{l-1},i_{l+1},\ldots,i_m) \in \mathrm{B}^{m-1}$ and we also denote the vector $\boldsymbol{i}$ by $(\boldsymbol{i}_{\overline{l}},i_l)$ by changing the order of elements. For any vector $\boldsymbol{x} \in \mathbf{R}^{\mathrm{B}^m \times J}$ and $l \in \{1,2,\ldots,m\}$, $x(\boldsymbol{i}_{\overline{l}},i_l;j)$ denotes the element $x(\boldsymbol{i};j)$ by changing the order of indices.

Let $(\boldsymbol{r}^1,\boldsymbol{r}^2,\ldots,\boldsymbol{r}^m;\boldsymbol{c})$ be a sequence of non-negative integer vectors where $\boldsymbol{r}^l \in \mathrm{Z}_+^{\mathrm{B}^{m-1} \times J}$ for each $l \in \{1,2,\ldots,m\}$ and $\boldsymbol{c} \in \mathrm{Z}_+^{\mathrm{B}^m}$. The element of $\boldsymbol{r}^l$ indexed by $(\boldsymbol{i};j) \in \mathrm{B}^{m-1} \times J$ is denoted by $r^l(\boldsymbol{i};j)$. The set of contingency tables corresponding to $(\boldsymbol{r}^1,\boldsymbol{r}^2,\ldots,\boldsymbol{r}^m;\boldsymbol{c})$ is defined by

$$\mathcal{T} \stackrel{\text{def.}}{=} \left\{ \boldsymbol{x} \in \mathrm{Z}_+^{\mathrm{B}^m \times J} \;\middle|\; \begin{array}{ll} x(\boldsymbol{i}_{\overline{l}},1;j) + x(\boldsymbol{i}_{\overline{l}},2;j) = r^l(\boldsymbol{i}_{\overline{l}};j) & (\forall l \in \{1,2,\ldots,m\},\ \forall \boldsymbol{i}_{\overline{l}} \in \mathrm{B}^{m-1},\ \forall j \in J), \\ \sum_{j \in J} x(\boldsymbol{i};j) = c(\boldsymbol{i}) & (\forall \boldsymbol{i} \in \mathrm{B}^m) \end{array} \right\}.$$

Each element in $\mathcal{T}$ is called a *table* for simplicity. In the following, $\sum_{\boldsymbol{i} \in \mathrm{B}^m} c(\boldsymbol{i})$ is denoted by N. Clearly, for any table $\boldsymbol{x} \in \mathcal{T}$, the sum total of elements of $\boldsymbol{x}$ is equivalent to N.

3 Markov Chain

Here, we propose a new Markov chain whose mixing time is bouded by a polynomial in n and $\ln N$. We define the parity function $\mathrm{p} : \mathrm{Z} \to \{1,-1\}$ by

$$\mathrm{p}(x) = \begin{cases} 1 & (x \text{ is an even integer }), \\ -1 & (x \text{ is an odd integer }). \end{cases}$$

For any index $\boldsymbol{i} \in \mathrm{B}^m$, we denote $\mathrm{p}(i_1 + i_2 + \cdots + i_m)$ by $\mathrm{p}(\boldsymbol{i})$. The vector $\Delta \in \{1,-1\}^{\mathrm{B}^m}$ is defined by $\Delta(\boldsymbol{i}) = \mathrm{p}(\boldsymbol{i})$ for each vector of indices $\boldsymbol{i} \in \mathrm{B}^m$. Given a pair of distinct column indices (j',j''), we define the vector $\Delta[j',j''] \in \mathrm{Z}^{\mathrm{B}^m \times J}$ by

$$\Delta[j',j''](j) \stackrel{\text{def.}}{=} \begin{cases} \mathbf{0} & (j \in J \setminus \{j',j''\}), \\ \Delta & (j = j'), \\ -\Delta & (j = j''). \end{cases}$$

For any table $\boldsymbol{x} \in \mathcal{T}$ and any pair of distinct column indices $\{j', j''\}$, we define the following set of vectors;

$$\mathcal{N}(\boldsymbol{x}; \{j', j''\}) \stackrel{\text{def.}}{=} \left\{ \boldsymbol{y} \in \mathrm{Z}_+^{\mathrm{B}^m \times \{j', j''\}} \;\middle|\; \begin{array}{l} x(i_{\overline{l}}, 1; j) + x(i_{\overline{l}}, 2; j) = y(i_{\overline{l}}, 1; j) + y(i_{\overline{l}}, 2; j) \\ \quad (\forall l \in \{1, 2, \ldots, m\},\ \forall i_{\overline{l}} \in B^{m-1},\ \forall j \in \{j', j''\}), \\ x(i; j') + x(i; j'') = y(i, j') + y(i; j'') \quad (\forall i \in B^m) \end{array} \right\}$$

$$= \left\{ \boldsymbol{y} \in \mathrm{Z}_+^{\mathrm{B}^m \times \{j', j''\}} \;\middle|\; \exists \theta \in \mathrm{Z},\ (\boldsymbol{y}(j'), \boldsymbol{y}(j'')) = (\boldsymbol{x}(j'), \boldsymbol{x}(j'')) + \theta(\Delta, -\Delta) \geq \boldsymbol{0} \right\}.$$

By using the above set $\mathcal{N}(\boldsymbol{x}; \{j', j''\})$, we propose our new Markov chain $\mathcal{M}^1$ with state space $\mathcal{T}$. For any table $\boldsymbol{x} \in \mathcal{T}$ and any pair of distinct column indices $\{j', j''\}$, we define the following set of tables;

$$\mathrm{N}^1(\boldsymbol{x}; \{j', j''\}) \stackrel{\text{def.}}{=} \{\boldsymbol{x}' \in \mathcal{T} \mid \boldsymbol{x}'(j) = \boldsymbol{x}(j)\ (\forall j \in J \setminus \{j', j''\}),\ (\boldsymbol{x}'(j'), \boldsymbol{x}'(j'')) \in \mathcal{N}(\boldsymbol{x}; \{j', j''\})\}$$
$$= \{\boldsymbol{x}' \in \mathcal{T} \mid \exists \theta \in \mathrm{Z},\ \boldsymbol{x}' = \boldsymbol{x} + \theta\, \Delta[j', j''] \geq \boldsymbol{0}\}.$$

Let $\mathcal{M}^1$ denote the Markov chain with the state space $\mathcal{T}$ with the following transition procedure. If X_t is the state of the chain $\mathcal{M}^1$ at time t and the element of X_t indexed by $(i; j)$ is denoted by $X_t(i; j)$. Then the state X_{t+1} at time $t+1$ is determined as follows. First, choose a pair of distinct column indices $\{j', j''\}$ randomly. Next, choose a table X_{t+1} from $\mathrm{N}^1(X_t; \{j', j''\})$ at random.

4 Mixing Time of New Markov Chain

The *mixing time* $\tau^1(\varepsilon)$ of $\mathcal{M}^1$ is defined by

$$\tau^1(\varepsilon) \stackrel{\text{def.}}{=} \max_{\boldsymbol{x} \in \mathcal{T}} \min\{t \mid \forall t' \geq t,\ \forall \mathcal{T}' \subseteq \mathcal{T}, -\varepsilon \leq \pi(\mathcal{T}') - \Pr[X_0 = \boldsymbol{x} \text{ and } X_{t'} \in \mathcal{T}'] \leq \varepsilon\ \},$$

where $\pi : \mathcal{T} \to [0, 1]$ is a unique stationary distribution of $\mathcal{M}^1$. To prove that our Markov chain is rapidly mixing, we use the path coupling method. We define a special Markov process with respect to $\mathcal{M}^1$ called coupling. A *coupling* of $\mathcal{M}^1$ is a Markov chain (X_t, Y_t) on $\mathcal{T} \times \mathcal{T}$ satisfying that each of $(X_t), (Y_t)$, considered marginally, is a faithful copy of the original Markov chain $\mathcal{M}^1$. More precisely, we require that

$$\Pr(X_{t+1} = \boldsymbol{x}' | (X_t, Y_t) = (\boldsymbol{x}, \boldsymbol{y})) = P_{\mathcal{M}^1}(\boldsymbol{x}, \boldsymbol{x}'),$$
$$\Pr(Y_{t+1} = \boldsymbol{y}' | (X_t, Y_t) = (\boldsymbol{x}, \boldsymbol{y})) = P_{\mathcal{M}^1}(\boldsymbol{y}, \boldsymbol{y}'),$$

for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}', \boldsymbol{y}' \in \mathcal{T}$ where $P_{\mathcal{M}^1}(\boldsymbol{x}, \boldsymbol{x}')$ and $P_{\mathcal{M}^1}(\boldsymbol{y}, \boldsymbol{y}')$ denote the transition probability from $\boldsymbol{x}$ to $\boldsymbol{x}'$ and from $\boldsymbol{y}$ to $\boldsymbol{y}'$ of the original Markov chain $\mathcal{M}^1$, respectively. The detail of our coupling is omitted. By using our coupling, it is known that we can analyse the mixing time of $\mathcal{M}^1$ (see [1]). Then the mixing time of our Markov chain $\mathcal{M}^1$ is as follows:

Theorem 1 *The Markov chain $\mathcal{M}^1$ is rapidly mixing with mixing time $\tau^1(\varepsilon)$ satisfying* $\tau^1(\varepsilon) \leq n(n-1)\ln(\lceil N/2^m \rceil \varepsilon^{-1})/2$. □

References

[1] R. Bubley and M. Dyer, Path coupling: A technique for proving rapid mixing in Markov chains, *38th Annual Symposium on Foundations of Computer Science*, IEEE, San Alimitos, 1997, pp. 223–231.

[2] P. Diaconis and L. Saloff-Coste, Random walk on contingency tables with fixed row and column sums, Technical Report, Department of Mathematics, Harvard University, 1995.

[3] M. Dyer and C. Greenhill, Polynomial-time counting and sampling of two-rowed contingency tables, *Theoretical Computer Sciences*, 246(2000), pp. 265–278.

Hardy-Weinberg 正確検定の p 値計算アルゴリズム

東大·情報理工 青木敏

1 はじめに

集団遺伝学の基礎原則として知られる Hardy-Weinberg の法則は，任意結婚と遺伝子頻度に関わる重要な理論である．Hardy-Weinberg 比率の検定としては古典的な適合度検定が用いられることが多いが，サンプル数が小さい場合やいくつかの遺伝子頻度が小さい場合には，漸近カイ二乗性の当てはまりの悪さが指摘されており，漸近論を用いない正確検定が必要とされる．Louis and Dempster (1987) は，allele 頻度を固定した上でのすべての遺伝子型頻度の数え上げによる正確検定のアルゴリズムを提案し，allele の数が 3 または 4 の場合の小さいサイズのデータに対して p 値の計算を行なったが，よりサイズの大きなデータに対しては，計算量の問題が生じる．本研究では，正確検定の p 値をより効率的に計算するアルゴリズムを提案する．

2 Hardy-Weinberg 則（平衡仮説）の正確検定

r 通りの allele, $A_1, A_2, ..., A_r$ を取るような多型遺伝子を考える．個々の観測値は A_iA_j の形の遺伝子型であり，頻度を上三角行列 $\boldsymbol{X}^o = (x^o_{ij})$, x^o_{ij} $(1 \le i \le j \le r)$ で表す．簡単のため，以後，$i > j$ に対しては $x_{ij} = x_{ji}$ と約束する．$\boldsymbol{y} = (y_1, y_2, ..., y_r)$, $y_i = x^o_{ii} + \sum_{j=1}^r x^o_{ij}$, $i = 1, ..., r$ と定義すれば，これは allele A_i の頻度であり，データのサイズを N とすれば，$\sum_{i \le j} x^o_{ij} = N$ および $\sum_{i=1}^r y_i = 2N$ が成り立つ．$\mathcal{F}$ を，allele 頻度が $\boldsymbol{X}^o$ と等しいような集合，

$$\mathcal{F} = \left\{ \boldsymbol{X} \mid \boldsymbol{X} = (x_{11}, x_{12}, x_{22}, ..., x_{rr}) \ge 0,\ x_{ii} + \sum_{j=1}^{r} x_{ij} = y_i \text{ for } i = 1, ..., r \right\}$$

と定義すれば，Hardy-Weinberg 則（平衡仮説）の下での $\boldsymbol{y}$ を与えたときの $\boldsymbol{X} \in \mathcal{F}$ の条件付き確率は，

$$P(\boldsymbol{X}) = \frac{N! \prod_{i=1}^r y_i!}{(2N)! \prod_{i \le j} x_{ij}!} 2^z, \quad z = \sum_{i<j} x_{ij} = N - \sum_{i=1}^{r} x_{ii}$$

で表される (Levene, 1949).

Hardy-Weinberg 則の条件付き検定の p 値は，

$$p = \sum_{\boldsymbol{X} \in \mathcal{T}} P(\boldsymbol{X}), \quad \mathcal{T} = \{ \boldsymbol{X} \mid \boldsymbol{X} \in \mathcal{F},\ P(\boldsymbol{X}) \le P(\boldsymbol{X}^o) \}$$

と定義される (Chapco, 1976 など)．これは，2×2 分割表に対する Fisher の正確検定（両側），あるいは，2 元分割表に対する Freeman-Halton 正確検定などの類似とも考えられる．

3 ネットワーク・アルゴリズム

この検定の p 値を計算するために，Louis and Dempster (1987) は，$\mathcal{F}$ の全ての要素を順次数え上げるアルゴリズムを提案しているが，N または r が極めて小さい例を除けば，計算量の問題が生ずる．本研究ではより効率的なアルゴリズムとして，ネットワーク・アルゴリズムを提案する．ネットワーク・アルゴリズムは，2元分割表の行と列の独立性の検定などの様々な設定で用いられているアルゴリズム (Mehta and Patel, 1983) であり，本研究では，同様のアルゴリズムがHardy-Weinberg正確検定にも構成でき，かつそれが有用であることを示す．

4 部分データに対する統計量の最大値，最小値の計算

提案するアルゴリズムでは，いくつかのalleleに対する遺伝子型頻度を順次条件付け，残りの部分に対する統計量の値の最大値，最小値を評価することにより，$\mathcal{F}$ の全ての要素を陽には数えずに，正確な p 値を計算することが可能となる．従って，与えられた $k(\leq r)$, $Y_1, Y_2, ..., Y_k$ に対して，次の整数計画問題を効率的に解くことが本質的に重要となる：

$$\text{minimize or maximize} \quad 2^z \left(\prod_{1 \leq i \leq j \leq k} x_{ij}! \right)^{-1}, \quad z = \sum_{1 \leq i < j \leq k} x_{ij},$$

$$\text{subject to} \quad x_{ij} + \sum_{j=1}^{k} x_{ij} = Y_i, \quad \text{for } i = 1, ..., k$$

$$x_{ji} = x_{ij}, \text{ かつ，非負の整数.}$$

本研究では，上の整数計画問題に対して，最小化問題の最適解の具体系と，最大化問題の最適解の二種類の上界評価の方法を与え，それらの値を用いたアルゴリズムによって，様々な例に対する正確検定の厳密な p 値が Louis and Dempster の方法の数倍～数十倍の効率で計算できることを確認した．

参考文献

[1] Chapco, W. (1976). "An exact test of the Hardy-Weinberg law", *Biometrics* **32**, 183-189.

[2] Levene, H. (1949). "On a matching problem arising in genetics", *Annals of Mathematical Statistics* **20**, 91-94.

[3] Louis, E. J., and Dempster, E. R. (1987). "An exact test for Hardy-Weinberg and multiple alleles", *Biometrics* **43**, 805-811.

[4] Mehta, C. R., and Patel, N. R. (1983). "A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables", *Journal of the American Statistical Association* **78**, 427-434.

簡単なアルゴリズムによる２項確率の計算

岡山理科大学・情報科学科：中村　忠
岡山大学・教育学部：平井　安久

はじめに

２項分布 B(n, p) の確率関数 $b(x; n, p)$ や分布関数 $B(x; n, p)$ は四則演算だけで計算可能な量である．現在，パソコンの市販ソフトでこれらの計算ができるものにはデータ処理のソフト(SPSS，S-Plus 等)や数式処理ソフト(マセマティカ等)がある．しかし，現在個人がパソコンの処理速度などの機能向上により，高価なソフトウェアや近似法を使わず直接計算できる環境が整っていると思われる．我々は２項分布の確率関数や分布関数の値を，直接，計算する簡単なアルゴリズムを開発した（この計算方法を“モグラたたき法”ということにする）．これまでに知られている方法で計算した値とモグラたたき法で計算した値の精度の比較をし，モグラたたき法と近似法との役割を論ずる．結果として，２項分布の確率関数や分布関数の値を標本の大きさが１億より小さいときはモグラたたき法で，それより大きいときは Peizer·Pratt の近似式で計算するのがよいことを提案する．

１．　２項確率の近似法

正規近似による方法としては，Stirling 公式を使用する方法，連続補正をした正規近似式，正規 Gram-Chalier 近似式(竹内(1975),竹内啓編(1997))，Camp-Paulson の近似式(Camp and Paulson,1951)，Peizer-Pratt の近似式(Peizer and Pratt,1968)などが知られている．これまでの数値結果によると，Peizer-Pratt の近似式IIが一番精度がよいことが報告されている．ここでは，我々が提案するモグラたたき法と Peizer-Pratt の近似式IIとの精度の比較をする．

２．　２項確率関数の計算アルゴリズム

２項確率関数 $b(x; n, p)$ の値を計算するアルゴリズムを述べる．ここで提案するアルゴリズムの特徴は次の３つである．①因数の積がオーバーフローを起こす可能性が少なくなるように積の順序の組み合わせを工夫する．②積の結果として算出される数がオーバーフローしないように，ある正数（ここでは 10^7 を採用）を越えたら，それより小さくなるように調整する．③積の結果として計算される数がアンダーフローしないように，ある正数（ここでは 10^{-7} を採用）より小さくなったら，それより大きくなるように調整する．このように上・下から数を調整する（叩く）ことから，このアルゴリズムを“モグラたたき法”と呼ぶことにする．

２項係数 nCx をつぎのような積の組み合わせに分解する．

$$nCx=\begin{cases}\prod_{i=1}^{x}(n-i+1)/x!, & 1\le x\le n/2,\\ \prod_{i=1}^{n-x}(n-i+1)/(n-x)!, & n/2\le x\le n.\end{cases}$$

例えば $1\le x\le n/2$ の場合は，$u=p\times q$ とおき，$p^xq^{n-x}=u^xq^{n-2x}$ の形にする（即ち $b(x; n, p)=(nCxu^x)q^{n-2x}$）．途中の積の計算結果ができるだけ大きくならないように積の順序を工夫する．$b(x; n, p)$ はつぎのような形の積になる．

$$b(x; n, p)=\begin{cases}\left[\prod_{i=1}^{m}\left(\frac{n-i+1}{x-i+1}u\times\frac{n-x+i}{i}u\right)\right]q^{n-2x}, & x=2m,\\ \left[\left[\prod_{i=1}^{m}\left(\frac{n-i+1}{x-i+1}u\times\frac{n-x+i}{i}u\right)\right]\left(\frac{n-m}{x-m}\right)\right]q^{n-2x}, & x=2m+1.\end{cases}$$

ここで，積$\prod_{i=1}^{m}\left(\frac{n-i+1}{x-i+1}u\times\frac{n-x+i}{i}u\right)$または$\left[\prod_{i=1}^{m}\left(\frac{n-i+1}{x-i+1}u\times\frac{n-x+i}{i}u\right)\right]\left(\frac{n-m}{x-m}\right)$の計算の順序は，$i=1,2,\cdots,x$の順におこなうものとする．各回において，それぞれの因数に対して10^7より大きいかどうかを調べる．10^7より大きい因数がある場合はその数が10^7以下になるまでqを繰り返し乗じる．ただし，これが可能であるためには因数qがあるとき，すなわち，$n-2x-k>0$でなければならない．ここにkはそれまでに使用されたqの回数である． もし，乗じるqがない場合はその数が10^7以下になるまで10^{-7}を繰り返し乗じる．もし，k回繰り返したなら，別に用意した何個かのバケツ（倍精度の配列で10の累乗の指数の値を格納するもの）に$7k$を加える．つぎに それぞれの因数に対して，10^{-7}より小さいかどうか調べる．10^{-7}より小さい因数がある場合はその数が10^{-7}以上になるまで10^7を繰り返し乗じる．もし，k回繰り返したなら，上で用意した何個かのバケツに$-7k$を加える．最後に，qのべき乗q^rが残っている場合($r>0$の場合)を考える．これに対しては1以上になるまでqに10を繰り返し乗ずる．k回繰り返したとして，$\tilde{q}=q\times10^k$とおく．$1\le\tilde{q}<10$は明らか．また，$-kr$を上で用意したバケツに加える．べき乗$\tilde{q}^r$の計算はよく知られている高速計算法を使用する．

3． 2項分布関数の計算アルゴリズム

前節で求めた確率関数の値を利用して，2項分布関数$B(x;n,p)=\sum_{i=0}^{x}b(i;n,p),\ \ x=0,\cdots,n$の値を求めるが，ここでは精度を落とさないために$\mathrm{B}(n,p)$のモード$M=[(n+1)p]$を基準にして場合分けする．例えば$x\ge M$の場合は，まず$b(M;n,p)$を前節のアルゴリズムにより求める，次に，残りの項については漸化式 $b(i;n,p)=\frac{n-i+1}{i}\frac{p}{q}b(i-1;n,p)$などを用いて，

$$B(x;n,p)=[\,b(M;n,p)+b(M+1;n,p)+\cdots+b(x;n,p)\,]$$
$$+[\,b(M-1;n,p)+b(M-2;n,p)+\cdots+b(0;n,p)\,].$$

と計算できる．ただし，大きなnに対しては(ア)膨大な項数の和の計算に時間がかかること，(イ)各計算を倍精度で行うので，その精度は高々15 ないし 16 桁であることの2点を考慮すれば，モードM（または平均np）からかなり離れたxに対する$b(x;n,p)$の値の小さい項の計算は省略可能となる．このようなxの範囲を見つけるために，以下のような改良されたUspenskyの不等式(Kambo & Kotz,1966)を用いる：

$$B(np-nc;n,p)<\exp\left(-\frac{nc^2}{2pq}-\frac{4nc^4}{9}\right),\quad p\le1/2\,.$$

これにより，計算精度を保持しながら計算時間を短縮することが可能になる．

4．数値実験による精度の比較

Cプログラムで作成したルーチンにより，2項分布の確率関数や分布関数の値をモグラたたきの方法で計算したものとPeizer-Prattの近似式で計算したものを比較した．その結果，標本の大きさが1億より小さいときはモグラたたき法で，それより大きいときはPeizer-Prattの近似式で計算するのがよいことがわかった．

NMRによる量子コンピュータの実現

岡山理科大学　工　水野陽一

我々は、これまでに位数発見問題の量子計算[1, 2, 3]と分割表上の量子ランダムウォークのシミュレーション[4]を行ってきたが、これと平行してNMR量子計算実験の研究を開始した。まず手始めとして、グローバーのファイル検索アルゴリズムを2キュービットの量子コンピュータ上で実行した。この量子コンピュータは、クロロホルム分子で構成されており、この分子中の水素原子と^{13}C炭素原子の2つの核スピンがキュービットの役割をしている。これらのキュービットの初期化、演算、読み出しの操作は、核磁気共鳴法を用いて実行される。

はじめに

NMRによる量子コンピュータの実験的研究は、有名なShorの因数分解量子アルゴリズム[5]が発見された4年後に、IBMのChuangらのグループ[6, 7]によって、世界に先駆けて行われた。その後Jonesらのグループ[8, 9]も実験に成功し、現在では2000年にChuangらが行った5キュービットの実験[10]が最高と思われる。今後、5キュービットを大きく超える実験が期待されるが、我々は手始めとして、クロロホルム分子を用いた2キュービットのNMR量子コンピュータ上でグローバーのファイル検索量子アルゴリズムを実行したので報告する。

グローバーの量子検索アルゴリズムの実行

集合$\{0,1,\ldots,n-1\}$の要素xについて$x=z$のとき$f(x)=1$で$x\neq z$のときは$f(x)=0$となるオラクル関数$f(x)$を考える。古典的にはこのオラクル関数を使ってzを発見するには$O(n)$回の試行が必要である。ところが、グローバーの量子アルゴリズム[11]を用いると$O(\sqrt{n})$回の試行で発見することができる。量子アルゴリズムでは、$f(x)$の代わりにx,y成分に$R_z(x,y)=(-1)^{f(x)}\delta_{xy}$を持つ選択的回転変換$R_z$を考える。すると平均に関する反転変換を$D=-WR_0W$とすると、ユニタリー変換$D_z=DR_z$をすべての状態が等しい確率をもつ状態$\phi_0=W\psi_0$に、およそ$\pi\sqrt{n}/4$回施すことによって解を見つけることができる。特に、2キュービットの場合は一回の試行で見つけることができる。2キュービットの$z=3$のときの量子コンピュータダイアグラムを図1に示す。

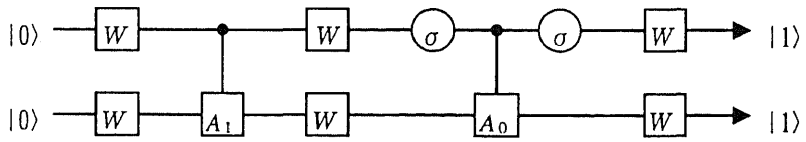


図1　$z=3$の場合のグローバーの量子コンピュータ

ただし、

$$A_0=\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}, A_1=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}$$

である。図に示した演算U_3Wを$\psi_0=|00\rangle$状態に施した後の^{13}Cの信号およびスワップ演算後の^{13}C信号をそれぞれ図2と図3に示す。図からわかるように、状態$\psi_3=|11\rangle$が観測されており、一度の演算で解$z=3$が見つけられている。

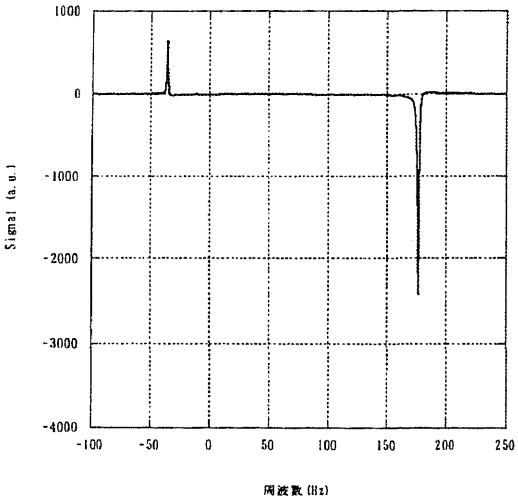


図2　演算 U_3W　後の ^{13}Cの信号

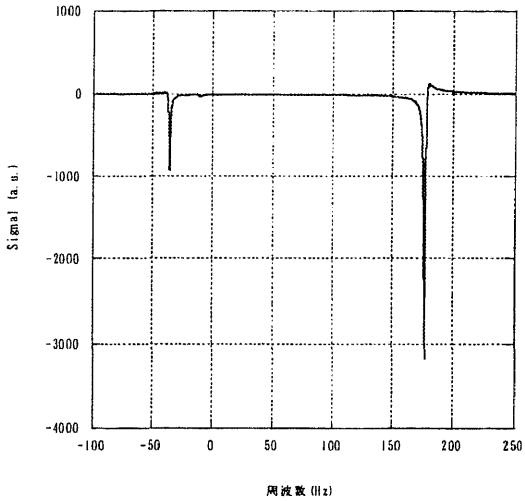


図3　スワップ演算後の ^{13}Cの信号

ディスカッション

実験結果から、一応2キュービットのNMR量子コンピュータ上でグローバーの量子アルゴリズムを実行する実験は成功したといってよい。しかし、逆算して得られる熱平衡状態の密度行列 ρ_{ex} は、かなりの実験誤差がある。これはスワップ演算後の信号強度が理論値に比べて3～4割程度小さいためである。この原因は、おもにパルス磁場の不均一によるものと考えられるが、他の演算に比べて誤差が大きく、パルスの組み合わせや、演算後のスピン温度が熱平衡状態から大きくずれていることも影響していると思われる。

今後、5キュービットを越える実験が期待されるが、最も障害になると思われるのは、キュービット数が増えるにしたがって分子が大きくなり、スピン間の距離が増大し、J 結合が弱くなることである。J 結合の強度は、1原子離れるごとに数分の1から10分の1程度に減少する。これは、直接的なキュービット間の演算は、直線的な分子では3～5キュービットが限界であることを示している。隣接間の相互作用のみで量子コンピュータを構成する方法も考案されているようで、高キュービットNMR量子コンピュータを実現するには、この方向の検討が必要ではなかろうか。

参考文献

1)R. Sawae, T. Sakata, Y. Matuda, K. Fukuda, M. Tei and K. Takarabe, The fourth Quantum Information Technology Symposium, Tokyo, Japan (2000)
2)R. Sawae, T. Sakata, Y. Matuda, K. Fukuda, M. Tei and K. Takarabe, in Proceeding of the first International Conference on Experimental Implementation of Quantum Computation, Sydney, Australia, to be appeared (2000)
3)R. Sawae, K Takarabe, T. Sakata, M. Tei, K. Fukuda and Y. Matuda, Technical Report of IEICE. p49-54 (2001)
4)R. Sawae, T. Sakata, K. Takarabe and M. Tei, The 5th Quantum Information Technology Symposium, Atugi, Japan (2001)
5)P.W. Shor, in Proceeding of 35th Annual Symposium on Foundation of Computer Science, IEEE Computer Society Press, Los Alamits, CA, pp.116 (1994)
6)I.L. Chuang, N. Gershenfeld and M. Kubinec, Phys. Rev. Lett. **80**, pp.3408-3411 (1998)
7)I.L. Chuang, L.M.K. Vandersypen, X. Zhou, D.W. Leung and S. Lloyd. Nature **393**, pp.143-146
8)J.A. Jones, M. Mosca and R.H. Hansen, Nature **393**, pp.344-346
9)J.A. Jones and M. Mosca, J. Chem. Phys. **109**, pp.1648-1653
10)L.M.K Vandersypen, M. Steffen, G. Breyta, C.S. Yannoi, R. Cleve and I.L. Chuang, Phys. Rev. Lett. **85**, pp.5452-5455 (2000)
11)L.K. Grover, Phys. Rev. Lett. **79**, p325 (1997)

量子通信とデコヒーレンスの研究

熊本大学・理　元吉明夫

1　序

量子力学の観測問題は未だに世界的な共通理解に達していない。Bennett et al. により提案された量子テレポーテーションは可能な新量子現象だが、ノイマンの射影仮説の言い替えである波束の収縮により生ずる EPR 相関に基づいている。我々は EPR 問題の正しい理解 [1,2,3] に基づき、色々の困難を生ずる射影仮説には依らない三重項の混合を補助に用いたテレポーテーションを提案した [2]。この立場は最近紹介され支持されている [4]。

もう一つの基礎的問題は「巨視的な量子系が如何に古典系に移行するか」であり、これにはデコヒーレンスが深く関係している。デコヒーレンスの詳しい解析は量子力学の基礎の確立の為にも、デコヒーレンスが障害となる量子情報処理にとっても重要である。

本報告では、我々のテレポーテーションの方法に基づき光子の偏光状態を用いて量子通信を実現する方法 [5] を述べる。

又、群 SO(3,1) の生成元から作った「粗視化された位置の作用素」を用いてデコヒーレンスの起源、量子系の古典系への移行 [6]、グリーン模型の修正 [7]、一般化されたマスター方程式による散逸過程 [8] を扱うことが出来ることも述べる。

2　量子通信

我々のテレポーテーションの方法における 3 粒子系の初期状態は

$$\begin{aligned}\hat{\rho}_{1,23} &= \frac{1}{4}\{|\psi\rangle_1\langle\psi|\otimes|1,1\rangle_{23}\langle 1,1| + |\psi\rangle_1\langle\psi|\otimes|1,-1\rangle_{23}\langle 1,-1| \\ &\quad +2|\phi\rangle_1\langle\phi|\otimes|1,0\rangle_{23}\langle 1,0|\}. \end{aligned} \tag{1}$$

である。ここで

$$|\phi\rangle_1 = a|x\rangle_1 + b|y\rangle_1, \quad |\psi\rangle_1 = -b^*|x\rangle_1 + a^*|y\rangle_1, \quad \langle\phi|\psi\rangle = 0, \quad |a|^2+|b|^2 = 1. \tag{2}$$

であり $|x\rangle \equiv |\leftrightarrow\rangle$, $|y\rangle \equiv |\updownarrow\rangle$ は光子の水平・垂直偏光で、状態 $|1,1\rangle$, $|1,0\rangle$, $|1,-1\rangle$ は三重項である。しかし、三重項の混合は光子では直接作ることが出来ず、光子の偏光は一光子づつしか測定出来ないことによりテレポーテーションは出来ないが、式 (1) の状態を $\hat{\rho}_{12,3}$ に組み替えて式の通りに装置を組むと量子通信が可能となる。

3　粗視化された位置の作用素

町田・並木の多ヒルベルト空間理論での分析から、測定は磁場等によるスペクトル分解過程と検出器による検出過程の 2 段階に分かれること、検出段階ではミクロの粒子の位置を測定するように全ての測定装置は出来ていること、実際にミクロの粒子の位置を測定する検出器局所系は粒子数不定の開放系であることが分かった。

この検出器の基本的性質はマクロのサイズ L をミクロの精度で決定することは事実上出来ないことと深く関係している。つまりマクロのサイズ L はミクロから見ると小さいが有限でランダムな ΔL の不定性（ゆらぎ）を持つと考えねばならない。このことの数学的表現の一つとして量子力学的位置の作用素

$$\hat{x}_j = i\hbar\{\frac{\partial}{\partial p_j} - \frac{\langle(\Delta L)^2\rangle}{\hbar^2}p_j\sum p_k\frac{\partial}{\partial p_k}\}, \tag{3}$$

を考えることができる。この作用素は射影座標を用いて運動量を

$$p_i \equiv \alpha_L^{-1/2} \cdot \xi_i/\xi_5, \tag{4}$$

と定義し、$\xi^2 = \xi_5^2 - \xi_1^2 - \xi_2^2 - \xi_3^2$ を不変にする群 SO(3, 1) の生成元から交換関係の解として作ることが出来る。$\Delta L \to 0$ はミクロの極限であり、マクロの極限は L に較べて ΔL を無視出来る極限として合理的に表現出来る。これから中心極限定理により $\langle(\Delta L)^2\rangle = N \cdot \langle(\Delta \ell)^2\rangle$ であり

$$\tilde{x}_j = i\hbar\{\frac{\partial}{\partial p_j} - \frac{\langle(\Delta\ell)^2\rangle}{\hbar^2} p_j \sum p_k \frac{\partial}{\partial p_k}\}, \quad \alpha_\ell \equiv \frac{\langle(\Delta\ell)^2\rangle}{\hbar^2}, \quad \alpha_\ell p^2 \equiv \gamma \ll 1, \tag{5}$$

を作ることが出来る。これを用いて、系のハミルトニアンが与えられていれば色々の問題を扱うことが出来る。先ず、作用素 (3) の固有値問題を解き、この作用素が適切な条件を備えていることを確かめた。

4 量子系から古典系への移行

簡単の為、マクロ系のモデルとして充分大きな数 N 個の原子からなる三次元結晶格子を考える。この系のハミルトニアンは調和近似で

$$H = \sum_{i=1}^{3N} \frac{p_i^2}{2m} + \sum_{ij} U_{ij} x_i x_j, \tag{6}$$

と書ける。これを基準座標に変換し、(5) の作用素を用いて、この系のエネルギー固有値を求めた。こうして系のエネルギーがミクロからマクロまで連続して繋がっている様子、ミクロでは離散的でマクロでは連続であり、レベルの識別が広範囲で出来なくなることを合理的に表現出来た。又、その波動関数を求め量子的振る舞いと古典的振る舞いを表現出来た。

同じ作用素を測定器のモデルとしての Green 模型に適用し、この作用素による修正で元の模型では起こらなかった波束の収縮を導くことが出来る。

同じ作用素を一般化されたマスター方程式に適用し散逸過程を導くことも出来る。その際の射影の物理的意味付けもはっきりし、巨視系は孤立系ではなく開放系と考えねばならないこと、それは内部に物理量のゆらぎを伴うこと、この作用素はこのゆらぎを適切に表現出来ていること等をはっきりさせた。

これらの研究により、この「粗視化された位置の作用素」は、量子系への極限も古典系への極限もうまく合理的に表現出来ることが分かる。

尚、この作用素は、もっと詳細な議論の出来るように一般化が可能であり、これが今後の課題である。

参考文献

[1] A. Motoyoshi, T. Ogura, K. Yamaguchi and T. Yoneda, Hadronic J. **20** (1997), 117.

[2] A. Motoyoshi, K. Yamaguchi, T. Ogura and T. Yoneda, P. T. P. **97** (1997), 819.

[3] S. Machida and A. Motoyoshi, Found. Phys. **28** (1998), 45.

[4] P Busch et al., Phys. Lett. A **284** (2001) 141.

[5] A. Motoyoshi and M. Matsuoka, Prog. Theor. Phys. **100** (1998) 455.

[6] T. Yoneda, A. Nagasato, T. Akamine, A. Motoyoshi, Nuovo Cim. B (2001) in print.

[7] T. Yoneda, A. Nagasato, T. Akamine, A. Motoyoshi, Nuovo Cim. B **116** (2001) 73.

[8] T. Yoneda, A. Nagasato, T. Akamine, A. Motoyoshi, Phys. Lett. A **280** (2001) 271.

確率的ニューラルネットワークとその応用

慶應大・理工 上辻茂男
慶應大・理工 柴田里程

1 確率的ニューラルネットワーク

確率的ニューラルネットワークは，各ニューロンがランダムに活性する，すなわち確率的ニューロンを含んだニューラルネットワークである．確率的ニューロンには，確率的に 0 か 1 の離散値を出力するニューロンと，ある確率分布に従う連続値を出力するニューロンの 2 つがある．これらの確率的ニューラルネットワークの特徴は，ある入力に対して常に同じ出力が得られるとは限らないことである．そのため，通常のニューラルネットワークとは異なり，予期せぬ動きに対して追従できる可能性を秘めている．また，学習法において通常のニューラルネットワークは，学習に用いる学習データの出力とネットワークから得られる出力との 2 乗誤差を最小にするパラメータを逐次推定するが，確率的ニューラルネットワークは，その尤度を定義し，パラメータを最尤推定する．しかし，全尤度でパラメータを一斉に推定することは，非効率であり計算的にも困難である．

そこで本報告では，離散型，連続型の確率的ニューラルネットワークの尤度をそれぞれ定義し，条件付尤度を導入し，モンテカルロ法を用いた簡単で効率的な学習法を提案している．

2 離散型と連続型の相違

訓練データ $(\mathbf{x}^{(i)}, \mathbf{t}^{(i)}), i = 1, \ldots, n$ として，確率的ニューラルネットワーク全体の対数尤度，

$$l = \sum_{i=1}^{n} \log P(\mathbf{t}^{(i)}|\mathbf{x}^{(i)})$$

を最大にするパラメータを各層ごとに定めることが目的である．しかし，離散型と連続型ではその学習法が異なる．

(1) 離散型の場合

今 K 層からなるニューラルネットワークを考える．第 1 層（最上層）が入力層，第 K 層（最下層）が出力層，それ以外の層を隠れ層とし，入力層以外のすべてのニューロンは確率的ニューロンであるとする．第 k 層からの M 次元出力ベクトルを $\mathbf{y}(k)$，第 k 層と第 $k+1$ 層の m 番目のニューロンとの間に荷せられた M 次元重みベクトルを $\mathbf{w}_m(k)$ とすると，第 $k+1$ 層の m 番目のニューロンが 1 を出力する確率は $f(\mathbf{w}_m(k)^T\mathbf{y}(k))$ と書ける．

第 k 層の出力 $\mathbf{y}(k)$ が与えられたときの第 $k+1$ 層の出力 $\mathbf{y}(k+1)$ の条件付確率は

$$P(\mathbf{y}(k+1)|\mathbf{y}(k)) = \prod_{m=1}^{M} \left\{f(\mathbf{w}_m(k)^T\mathbf{y}(k))\right\}^{y_m(k+1)} \left\{1 - f(\mathbf{w}_m(k)^T\mathbf{y}(k))\right\}^{1-y_m(k+1)}$$

とかける．今，第 k 層に注目して，$E^{\mathbf{y}(k)|\mathbf{x}^{(i)}}$ を入力 $\mathbf{x}^{(i)}$ が与えられたときの $\mathbf{y}(k)$ での期待値とすれば，全尤度 l を

$$l = \sum_{i=1}^{n} \log \sum_{\mathbf{y}(k)\in\{0,1\}^M} P(\mathbf{t}^{(i)}|\mathbf{y}(k))P(\mathbf{y}(k)|\mathbf{x}^{(i)}) = \sum_{i=1}^{n} \log E^{\mathbf{y}(k)|\mathbf{x}^{(i)}} \left[P(\mathbf{t}^{(i)}|\mathbf{y}(k))\right]$$

と表すことができる．ここで，$\sum_{\mathbf{y}(k)\in\{0,1\}^M}$ は M 個の 0 か 1 のすべての組み合わせに関する和，$E^{\mathbf{y}(k)|\mathbf{x}^{(i)}}$ は，入力 $\mathbf{x}^{(i)}$ が与えられたときの $\mathbf{y}(k)$ での期待値である．これは第 k 層から下の層全体と，それ以外の上の層の 2 つにネットワークに分割して全尤度を表したものである．. そして，$P(\mathbf{t}^{(i)}|\mathbf{y}(k))$ がその第 k 層から下のネットワーク全体の条件付尤度である．ニュートン法で各層のパラメータを最尤推定する．その際，モンテカルロ法を用いて期待値を評価すればより効率的に計算ができることが分かった．また，この方法は，常に全尤度を最大にするパラメータを求めている．したがって，条件付尤度の導入は計算的困難を避けるためであると考察できる．

(2) 連続型の場合

各ニューロンは正規分布にしたがう値を出力すると考える．全体のモデルは，

$$\mathbf{y}(K) = W(K-1)^T W(K-2)^T \cdots W(1)^T \mathbf{x} + \sum_{j=2}^{K} W(K)^T W(K-1)^T \cdots W(j)^T \boldsymbol{\epsilon}(j)$$

である．ただし，$W(k) = (\mathbf{w}_1(k), \ldots, \mathbf{w}_M(k))$ である．今，全尤度に基づいて最尤法を行うと，各層のパラメータの積，$W(K-1)^T \cdots W(1)^T$ しか求めることができない．したがって，各 $W(k)$ は同定不可能となる．しかも，離散型のように条件付尤度を用いて，全尤度を分割するこはできない．そこで，条件付尤度の最大化も要求する．条件付尤度は，

$$\prod_{i=1}^{n} f(\mathbf{t}^{(i)}|\mathbf{y}(k)^{(i)})$$

とする．ただし，$f(\mathbf{t}|\mathbf{y}(k))$ は第 k 層の出力 $\mathbf{y}(k)$ を与えたときの $\mathbf{t}$ の条件付密度であり，$\mathbf{y}(k)^{(i)}$ は i 番目の訓練データの入力 $\mathbf{x}^{(i)}$ が与えられたときの第 k 層の出力である．今，$X = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})^T$, $Y = (\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(n)})^T$ として，全尤度を最大にするパラメータの積は，

$$W(1)\cdots W(K-1) = (X^T X)^{-1} X^T Y$$

と求まり，それを制約条件とし，各層での条件付尤度の最大化を解き，$W(k)\cdots W(K-1)$ を推定する．そして，各 $W(k)$ を同定していく．この際，先ほど同様，モンテカルロ法で計算を簡略化することができた．したがって，条件付尤度の導入は，同定不可能な問題を同定可能とするために導入していると考察できる．

また，離散型と連続型の場合で，能率的にパラメータの推定を行うアルゴリズムの構築，また，そのアルゴリズムの収束性も証明できた．

確率的ニューラルネットワークは，確率的な動きを表現するランダムな要素を持ち合わせた学習モデルであり，その収束性も保証されていることからデータ解析の有益な道具と成り得ると考えられる．今後は，データの背後の現象の説明に適しているかどうか入念なデータ解析を行った上で，データ解析の道具として有効であることを検証したい．

ニューラルネットワークと層別因子を含む線形回帰分析との数値実験による比較

大東文化大学法学部　　浅野美代子

1．はじめに

ニューラルネットワークを回帰分析としてデータ解析を行う場合の数値実験を行った（浅野(2001）．ニューラルネットワークの特徴を顕著に表す２つの数値実験である．ここでは，数値実験結果より導かれた事項のひとつである構造変化（変化点問題）を，ニューラルネットワークで解析することが有効であることを実データ解析によって示した．通常のノンパラメトリックな回帰分析がデータの平滑化作業に基づくものに対し，ニューラルネットワークモデルは変化点と変化率を与えてｓ字曲線の重ね合わせで回帰関数を近似することが大きな特徴である．このため，中間層の出力を観察することで変化点を効果的に抽出することができる．変化点問題は，数理統計的には，尤度あるいは Bayse 的接近が，Hinkley(1996), Hinkley and Hinkley(1970), Hinkley and Schechtman(1987), Barry and Hartigan(1993)によって，累積和技法を用いた方法については，仁科（1986）などの研究がある．実際，変化点の位置，個数が未知の場合に，その構造を線形モデルで表現しようとすると，データ系列の数だけ層別因子（ダミー変数）を，説明変数に加える必要があり，その変数群からモデル選択を行い変化点の場所や個数を探索することが考えられる．しかし，情報量規準を用いたとしても探索の多重性の影響をどのように調整するかは難しい．

2．数値例

国内総支出 GDE(=GDP)について，ニューラルネットワークの入力変数としては，補助変数時間(t, t=1, 2, ･･･)１変数で解析した結果を表１に示す．AIC 最小は，中間層ユニット数が３であった．この解析における中間層の各ユニット１から３の出力値にウエイトを乗じた分解図（Sigmoid decomposition）と，GDP 実績，予測値のグラフを図１に示した．各ユニットの意味付けは、考察が必要であるが，この結果から，構造変化の変化点をみつけて層別変数を１つ含む回帰分析を行った．２水準のダミー変数を z11，３水準のダミー変数を z21，z 22，４水準のダミー変数を z31，z 32，z 33，５水準のダミー変数を z41，z 42，z 43，z44 と設定した．吉澤（1992）第９章を参考にして，時間とダミー変数の積の項を説明変数に加えて解析した回帰分析結果を表２にまとめた．表１と表２の

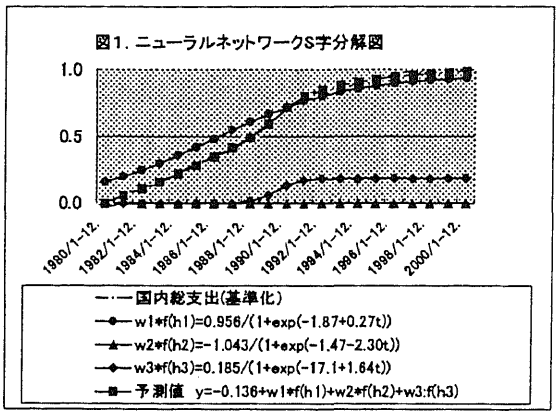
表1．ニューラルネットワーク

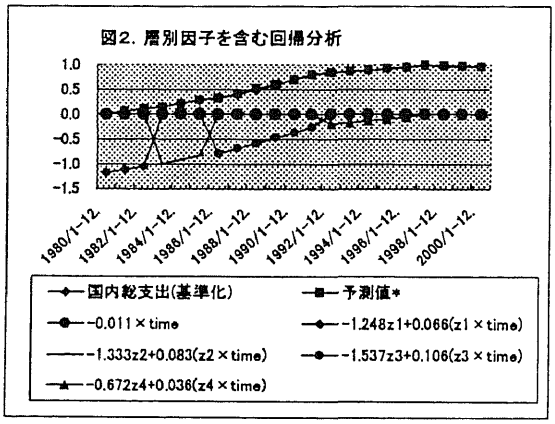
入力変数	中間層のユニット数	ＲＳＳ	ＡＩＣ
t	1	0.01086	−152.9
t	2	0.00745	−154.8
t	3	0.00379	−163.0
t	4	0.00307	−161.5

表2．線形回帰分析

水準	説明変数	重相関係数	ＲＳＳ	ＡＩＣ
0	t	0.9698	0.14995	−99.8
	t, t×t	0.9886	0.05732	−118.0
2	t, z11	0.9810	0.09477	−107.4
	t, z11, t×z11	0.9965	0.01771	−140.6
3	t, z21, z22	0.9889	0.05537	−116.7
	t, z21, z22, t×z21, t×z22	0.9933	0.03351	−123.3
4	t, z31, z32, z33	0.9811	0.09429	−103.5
	t, z31, z32, z33, t×z31, t×z32, t×z33	0.9987	0.00671	−153.0
5	t, z41, z42, z43, z44	0.9877	0.06146	−110.5
	t, z41, z42, z43, z44, t×z41, t×z42, t×z43, t×z44	0.9996	0.00211	−173.3

AICを比較すると，水準4の層別因子，時間と，ダミー変数との積の項を含む回帰分析が一番良い結果であった．図2に，この解析結果を，補助変数時間(Time)と4つのダミー変数(z1，z2，z3，z4)ごとの分解図で示した．





*)予測値＝1.196-0.011×time-1.248z1+0.066(z1×time)-1.333z2+0.083(z2×time)
-1.537z3+0.106(z3×time)-0.672z4+0.036(z4×time)

3．結論

GDPを例にとって，変化点構造をニューラルネットワークで解析し，変化点を把握し利用する方法を示した．ニューラルネットワークのもつこれらの性格は，地点での変化を有する問題などに対しても有効と考えられる．

参考文献．

Asano.M., Tsubaki,H., Yoshizawa,T.(2001) *Effectiveness of neural networks to regression with structural changes*, The special issue in Journal of Applied Stochastic Models in Business and Industry. 投稿中.

Hinkley,D.V.(1969). *Inference about intersection in two-phase regression,* Biometrika56, 495-504

Hinkley, D.V. and Hinkley, E.A. (1970). *Inference about the change point in a sequence of binominal random variables,* Biometrika, 57, 477-488.

Muller,H.G.(1992). *Change-points in nonparametric regression analysis,* Annals of Statistics, Vol.20, No.2, 737-761.

Nishina.K. (1986) . *Estimation of the Change-point from Cumulative Sum Tests,* Rep. Stat. Appl. Res., JUSE, Vol.33, No.4, 1-14.

浅野美代子(2001)．ニューラルネットワークを用いた層別因子を含む回帰構造の解析，計算機統計学，投稿中．

伊藤博・市川慶良・山内修・浅野良晴・馬場宗・浅野美代子（1999)．事務所ビルにおける瞬時流量算出法に関する研究．水道協会雑誌，第766号，28-36.

吉澤正（1992)．統計処理．岩波書店．

http://www.esri.cao.go.jp/jp/sna/qe011/gdemenuj.html.

GDP，出典 ： 経済社会総合研究所，平成7暦年基準GDE（GDP）需要項目別時系列表，

大東文化大学法学部 〒175-8571 東京都板橋区高島平1-9-1 E-mail asano@ic.daito.ac.jp

大量データの視覚化と最尤多次元尺度構成法

宿久　洋 (鹿児島大学 理学部), 橋口　博樹, 岡　隆一 (新情報処理開発機構)

1　はじめに

本報告では, 大量データの記述のための多次元尺度構成法 (MDS) の提案を行う. 大量データからそのまま非類似性行列を作成し MDS を利用する場合, その結果は複雑なものとなり, 解釈が困難な場合が多い. この問題の 1 つの解決法として, k-means 等のクラスタイング法により予めグループ分けをし, 各グループの代表値 (例えば, 平均) をデータとして MDS を適用することが考えられる. この場合, グループ数がデータ数となり結果の解釈は容易になる. しかしながら, この方法が多くの情報の損失を伴うのは指摘するまでもない. そこで我々は, 単に代表値から求められた結果のみを表示するのではなく元データの持つ情報も合わせて同一の空間に表示することを提案する. 具体的には, 元データを何らかの手法でグループ分けし, 異なる 2 つのグループに属する対象間の非類似性をそのグループ間の非類似性の繰り返し観測だと考え, Ramsay (1977, 1978, 1982) で提案されている最尤多次元尺度構成法 (MLMDS) を用いて対象の付置及び漸近信頼区域を求めることを提案する.

2　クラスター化と付置座標および漸近信頼区域の決定

N 個のデータ $\boldsymbol{y}_\alpha$ $(\alpha = 1, \ldots, N)$ が k-means 等の何らかのクラスタリング法により n 個のクラスター C_i $(i = 1, \ldots, n)$ に分類されているとする. もちろん, 外的な情報により分類されていてもかまわない. この時, クラスター C_i と C_j の間の非類似性に関して, R_iR_j 個の

$$d_{ijr_ir_j} = \left[(\boldsymbol{y}_{r_i} - \boldsymbol{y}_{r_j})'(\boldsymbol{y}_{r_i} - \boldsymbol{y}_{r_j})\right]^{1/2} \tag{1}$$

を計算する. ここで, $\boldsymbol{y}_{r_i}$, $\boldsymbol{y}_{r_j}$ はそれぞれ C_i, C_j に属する r_i 番目, r_j 番目の対象のデータを表している. この R_iR_j 個の $d_{ijr_ir_j}$ に適当に順番をつけた $d_{ijr_{ij}}$ $(r_{ij} = 1, \ldots, R_iR_j)$ を 2 つのクラスター間の非類似性の観測値として扱うことにする.

最も簡単な MDS の利用法はこの観測値の平均

$$\bar{d}_{ijr_{ij}} = \sum_{r_{ij}=1}^{R_iR_j} d_{ijr_{ij}}/R_iR_j \tag{2}$$

を非類似性データと考えることであろう. この時, 付置を求めるべき対象の個数はクラスター数 n 個となり, 計算も解釈も容易になる. しかし, 当然のことながら元データに対して利用するデータが少なくなったのに対応した容易さであり満足できる結果とは言い難い.

この二律背反の問題への 1 つの解決策として, Ramsay(1977) の MLMDS による対象の付置の決定および Ramsay (1978) に基づく付置の信頼区域の決定を提案する.

2.1　MLMDS による付置の決定

$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ $(i = 1, \ldots, n)$ を推定すべき n 個の対象の k 次元空間における座標とし, その対象間距離を $d^*_{ij} = \left(\sum_{m=1}^{k}(x_{im} - x_{jm})^2\right)^{1/2}$ $(i, j = 1, \ldots, n)$ とする. $d_{ijr_{ij}}$ を d^*_{ij} の r_{ij} 番目の観測とし, これが i. i. d. で密度関数 $f(d_{ijr_{ij}})$ を持つ分布に従っていると仮定する. ここでは, 対数正規モデル: $\log d_{ijr_{ij}} \sim N(\log d^*_{ij}, \sigma^2)$ および正規モデル: $d_{ijr_{ij}} \sim N(d^*_{ij}, (\sigma d^*)^2)$ の

2つのモデルを考え, Ramsay (1977) に従い, 最尤法を用いて付置 $\boldsymbol{x}_i$ と散布度パラメータ σ^2 を推定する. 紙面の都合により詳細は省略する.

2.2 付置の漸近信頼区域の決定

最尤多次元尺度法の利点は, 与えられたデータをすべて用いたクラスターの付置を決定できるだけでなく, 最尤法を用いていることから, 適合度検定や漸近信頼区域の推定も可能となることである. ここでは, Ramsay (1978) で提案された方法に基づき, 漸近信頼区域の推定について述べる. 以下, $\boldsymbol{\theta} = \mathrm{Vec}(\boldsymbol{X})$ すなわち, $\boldsymbol{X}$ の要素を適当に並べた $1 \times nk$ ベクトルを扱う.

パラメータの推定に関して, 多次元尺度構成法の場合は解析結果に回転や拡大縮小の自由度があるため, 何らかの制限を加えない限り, 一意に定まらない. このような制約 $q(\boldsymbol{\theta})$ について, Ramsay (1978) では, 回転や拡大縮小の自由度を抑えるという制約に加えて, $q(\boldsymbol{\theta})$ が $\boldsymbol{\theta}$ の最尤推定量で最大値をとる関数という仮定をおき, $q(\boldsymbol{\theta})$ と $\log L(\boldsymbol{\theta})$ の両方を最大にする $\boldsymbol{\theta}^*$ が (一意に) 存在すると仮定している. また, このような $\boldsymbol{\theta}^*$ を $\boldsymbol{\theta}$ の制約つき最尤推定量と呼んでいる. この制約つき最尤推定量 $\boldsymbol{\theta}^*$ は,

$$g(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) + q(\boldsymbol{\theta}) \tag{3}$$

を最大にするので, これを具体的に求める際には, (3) を最大化しながら解を得る最急降下法などを用いる. なお, 他のパラメータ σ^2 については前節と同様に求める.

漸近信頼区域については, Ramsay(1978) により,

$$\Pr[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \mathrm{Cov}^{-1}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq \chi^2_{p,\alpha}] = 1 - \alpha \tag{4}$$

が導かれているので, 以下を満たす $\boldsymbol{\theta}$ の変域として求めることができる.

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \mathrm{Cov}^{-1}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq \chi^2_{p,\alpha} \tag{5}$$

ただし,

$$\boldsymbol{T}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}) - \boldsymbol{Q}(\boldsymbol{\theta}) \tag{6}$$

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}\left[\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)'\right] = -\mathrm{E}\left[\frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \tag{7}$$

$$\boldsymbol{Q}(\boldsymbol{\theta}) = \frac{\partial^2 q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \tag{8}$$

であり, $\chi^2_{p,\alpha}$ は自由度 p のカイ二乗分布の $100\alpha\%$ 点である. また, p は $\mathrm{Cov}^{-1}(\hat{\boldsymbol{\theta}})$ の階数である.

多次元尺度構成法の場合はパラメータ全体の同時信頼区域ではなく, 各対象を規定している n 個の座標を同時に考慮した点ごとの信頼区域に興味がある. $\hat{\boldsymbol{\theta}}$ の漸近正規性から, その一部のパラメータの周辺分布も正規分布に従い, その分散共分散行列の推定値も $\mathrm{Cov}(\hat{\boldsymbol{\theta}})$ のそのパラメータに対応する部分行列 $\mathrm{Cov}(\hat{\boldsymbol{x}}_i)$ として与えられる. よって, 対象ごとの周辺信頼区域も

$$(\hat{\boldsymbol{x}}_i - \boldsymbol{x})' \mathrm{Cov}^{-1}(\hat{\boldsymbol{x}}_i)(\hat{\boldsymbol{x}}_i - \boldsymbol{x}) \leq \chi^2_{k,\alpha} \tag{9}$$

を満たす $\boldsymbol{x}$ の変域として求めることができる.

参考文献

[1] Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, **42**, 241–266.

[2] Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika*, **43**, 145–160.

[3] Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data (with discussion) *Journal of Royal Statistical Society, Series A*, **145**, 285–312.

層別逆回帰モデルの数理的考察

北海道大学　水田 正弘

1　はじめに

層別逆回帰 (Sliced Inverse Regression; SIR) は、Li, Ker-Chau[1] により提案された手法であり、回帰分析における説明変数空間の次元縮小を目的にしたものである。層別逆回帰で利用されているモデルは非常に一般的なものであり、汎用性が高い。具体的なアルゴリズムとしては、SIR1、SIR2、拡張 SIR、射影追跡による SIR などがある。これらの手法の一部については、数値例を中心として、その長所や欠点が議論されている。

本論文では、層別逆回帰で仮定されているモデル (以下、SIR モデルと呼ぶ) について理論的な考察を行う。

2　層別逆回帰のモデルとアルゴリズム

回帰分析において説明変数の個数が多い場合には、回帰係数などのパラメータの分散が大きくなりやすいなどの問題が起こる。そこで、(説明) 変数選択や説明変数の空間を縮小する方法が多数提案されている。以下では、説明変数の空間を縮小する方法の 1 つである層別逆回帰について述べる。

SIR では、モデルとして、

$$y = f(\boldsymbol{\beta}_1 \boldsymbol{x}, \boldsymbol{\beta}_2 \boldsymbol{x}, \cdots, \boldsymbol{\beta}_K \boldsymbol{x}, \varepsilon) \tag{1}$$

を想定する。ここで、$\boldsymbol{x}$ は p 次元の説明変量を表す列ベクトル、$\boldsymbol{\beta}_k$ は未知の行ベクトル、ε は $\boldsymbol{x}$ と独立な確率変数、f は $\boldsymbol{R}^{K+1}$ 上の任意な未知の関数である。すなわち、p 次元の説明変数の K 次元部分空間だけで, y を知るために必要なことを全て入手できると仮定したモデルである。

SIR における目的は、関数 f を求めることではなく、K 次元部分空間すなわちベクトル $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K$ を求めることである。この $\{\boldsymbol{\beta}_i\}(i = 1, 2, \cdots, K)$ を有効次元縮小方向 (effective dimension reduction directions; e.d.r. 方向)、これらのベクトルにより張られる K 次元部分空間を有効次元縮小空間 (e.d.r. 空間) とよぶ。

3　条件付き分布の考察

$E[\boldsymbol{X}|y]$ だけでは、e.d.r. 方向を見つけるには不十分である。そこで、条件付き分布 $\boldsymbol{X}|y$ を直接、検討する。以下では、$\boldsymbol{x}$ は標準正規分布に従うと仮定する。つまり、$\boldsymbol{x} \sim N(0, I_p)$ とする。もし、一般の正規分布であるなら、アフィン変換により、$\boldsymbol{x}$ を球化することができる。一般性を失わず、$\boldsymbol{\beta}_i \boldsymbol{\beta}_j^T = \delta_{ij}, (i, j = 1, 2, \cdots, K)$ と仮定する。そこで $\{\boldsymbol{\beta}_i\}$ $(i = 1, 2, \cdots, p)$ が $\boldsymbol{R}^p$ における正規直交系となるような $\boldsymbol{\beta}_i (i = K+1, \cdots, p)$ を選ぶことができる。

$\boldsymbol{x}$ は、$N(0, I_p)$ に従うので、$(\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_p\boldsymbol{x})$ も同様に $N(0, I_p)$ に従う。簡単のために $x_i = \boldsymbol{\beta}_i\boldsymbol{x}$ とおく。変数変換 ϕ を以下のように定義する。

$$\phi(x_1, \cdots, x_p, \varepsilon) = (x_1, \cdots, x_p, y).$$

ヤコビアン J^{-1} は

$$J^{-1} = \frac{\partial(x_1, \cdots, x_p, y)}{\partial(x_1, \cdots, x_p, \varepsilon)} = \left| \frac{\partial y}{\partial \varepsilon} \right|$$

となる。ここで、$\frac{\partial y}{\partial \varepsilon}$ は、$x_1, \cdots, x_k$ の関数である。そこで、$(\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_p\boldsymbol{x}, y)$ の密度関数は、

$$h(\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_p\boldsymbol{x}, y) = \phi(\boldsymbol{\beta}_1\boldsymbol{x}) \cdots \phi(\boldsymbol{\beta}_p\boldsymbol{x}) \psi(\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K\boldsymbol{x}, y),$$

となることは容易に分かる。ここで、$\phi(x) = 1/\sqrt{2\pi} \exp(-x^2/2)$ とし、$\psi(\cdot)$ は $\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K\boldsymbol{x}, y$ の関数である。

条件付き密度関数は、

$$h(\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_p\boldsymbol{x} \mid y) = \phi(\boldsymbol{\beta}_{K+1}\boldsymbol{x}) \cdots \phi(\boldsymbol{\beta}_p\boldsymbol{x}) g(\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K\boldsymbol{x}),$$

となる。ただし、$g(\cdot)$ は、$\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K\boldsymbol{x}$ の関数であり、通常、正規分布の密度関数とはならない。以上により、$h(\boldsymbol{\beta}_1\boldsymbol{x}, \cdots, \boldsymbol{\beta}_p\boldsymbol{x} \mid y)$ は、正規分布 $\phi(\boldsymbol{\beta}_{K+1}\boldsymbol{x}) \cdots \phi(\boldsymbol{\beta}_p\boldsymbol{x})$ と非正規分布 $g(\cdot)$ に分解することができる。

射影追跡は、非線形な構造を見つけ出すために開発された手法である。特に、Friedman (1987) は、非線形な構造である規準として、非正規性を利用している。SIRpp (Mizuta; 1999) は、この考え方に基づいている。

4 おわりに

本報告では、SIR の考え方を紹介するとともに、射影追跡法を SIR に利用する方法の提案と層別逆回帰モデルの考察結果を報告した。SIR の特徴の一つとして、計算量が少ないことが上げられる。それに対して、提案手法ではスライスの個数 H だけ射影追跡を実行しなくてはいけない。しかし、一般にスライスの個数は、それほど多くする必要がないので、提案手法は十分実用的な計算時間で実行可能である。しかし、SIR や SIR II との詳細な比較、スライスの方法、スライス数の設定、重み $w(h)$ の設定法、射影追跡法の選択など、多くの課題が残されている。

参考文献

[1] Li, Ker-Chau(1991). Sliced Inverse Regression for Dimension Reduction. *JASA* Vol.86 No.411, 316–342.

[2] Mizuta, M. (1999). Sliced Inverse Regression with Projection Pursuit, In H. Bacelar-Nicolau, F. Costa Nicolau and J. Janssen (Eds.): *Applied Stochastic Models and Data Analysis*. INSTITUTO NACIONAL DE ESTATÍSTICA), 51–56.

[3] Mizuta, M. and Minami. H.(2000). An Algorithm with Projection Pursuit for Sliced Inverse Regression Model, *Data Analysis, Classification, and Related Methods*, (H.A.L.Kiers, J.-P.Rasson, P.J.F.Groenen and M.Schader Eds.) Springer, 255-260.

[4] Friedman, J. H. (1987). Exploratory Projection Pursuit. *Journal of the American Statistical Association, 82*, 249–266.

反復ブートストラップ法とその近似

九州大学大学院経済学研究院　前園宜彦

元の標本から得られた再抽出標本から，さらにランダムに抽出して推測の精度を上げていくブートストラップ反復法の近似について考察する．

$\mathcal{X}=\{X_1,\ldots,X_n\}$ を元のデータとし，$\mathcal{X}^*=\{X_1^*,\ldots,X_n^*\}$ を $\mathcal{X}$ から復元抽出でランダムに得られた再抽出標本，さらに $\mathcal{X}^{**}=\{X_1^{**},\ldots,X_n^{**}\}$ を $\mathcal{X}^*$ から復元抽出でランダムに得られた再抽出標本とする．このときダブル・ブートストラップ法は $\mathcal{X},\mathcal{X}^*,\mathcal{X}^{**}$ に基づいて推測を行なうもので，推測の精度は上がるが，冪のオーダーで計算量が増えるという問題点がある．この問題の解決のために様々な改良が提案されている．ここではブートストラップ反復法に対して，計算量のオーダーを下げるが，推測の精度をあまり下げない手法を提案し，シュミレーションでその有効性を吟味した．$\mathcal{X}^*$ が与えられたときに各 X_i の重み

$$p_i^*=n^{-1}\,(\text{number of appearances of } X_i \text{ in } \mathcal{X}^*)$$

を考える．$p^*=(p_1^*,\ldots,p_n^*)$ とおき $\hat{a}^*=\hat{f}(p^*)$ を推測に使う量とする．$p^0=(n^{-1},\ldots,n^{-1})$ を通常のリサンプリングを行なうときの一様分布とし，$\hat{a}=\hat{f}(p^0)$ を $\hat{a}^*$ の通常のブートストラップに基づく推測量とする．$\hat{f}$ をなめらかな関数とすると

$$\hat{f}(p)=\hat{f}(p^0)+\sum_{i=1}^{n}(p_i-p_i^0)\hat{f}_i(p^0)+\frac{1}{2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}(p_{i_1}-p_{i_1}^0)(p_{i_2}-p_{i_2}^0)\hat{f}_{i_1i_2}(p^0)+\ldots$$

と展開される．ここで $\hat{f}_{i_1\ldots i_r}(p)$ は $(\partial^r/\partial p_{i_1}\ldots\partial p_{i_r})\,\hat{f}(p)$ または偏差の適当な近似とする．

ここで Calibration 法による信頼区間の構成に関する Weighted-bootstrap によるアプローチについて考えよう．$\hat{\theta}$ を θ の推定量とし，$\hat{\theta}^*$ を $\mathcal{X}^*$ に基づく対応する推定量とする．さらに $T=n^{1/2}(\hat{\theta}-\theta)$，$T^*=n^{1/2}(\hat{\theta}^*-\hat{\theta})$ で T^{**} を T^* に対応する量とする．$\hat{x}(\alpha)$ と $\hat{x}^*(\alpha)$ を $P\{T^*\le\hat{x}(\alpha)|\mathcal{X}\}=\alpha$ 及び $P\{T^{**}\le\hat{x}^*(\alpha)|\mathcal{X}^*\}=\alpha$ とする．$\beta=\hat{\beta}(\alpha)$ を

$$P\{T^*\le\hat{x}^*(\beta)|\mathcal{X}\}=\alpha$$

の解とする．ダブル・ブートストラップ Calibration である $\hat{\beta}$ を使うと被覆確率の改善がなされ，信頼区間 $\hat{\mathcal{I}}(\alpha)=(-\infty,\hat{\theta}-n^{-1/2}\hat{x}\{\hat{\beta}(1-\alpha)\})$ は $\alpha+O(n^{-1})$ の被覆確率を持つ．これは

$$P[T\le\hat{x}\{\hat{\beta}(\alpha)\}]-\alpha=O(n^{-1})$$

から導かれる．このとき $\hat{x}(\alpha)=\hat{f}(p^0)$ 及び $\hat{x}^*(\alpha)=\hat{f}(p^*)$ が $\hat{a}$, $\hat{a}^*$ の例である．

本報告で提案する Weighted-bootstrap による $\hat{\beta}$ の近似 $\tilde{\beta}$ は $\tilde{\beta}(\alpha) = \hat{Q}^{-1}(\alpha)$ で与えられる．ここで

$$\hat{Q}(\beta) = P\left\{T^* \leq \hat{x}(\beta) + \sum_{i=1}^{n} (p_i^* - p_i^0)\, \hat{f}_i(p^0) \middle| \mathcal{X}\right\}.$$

$\hat{\beta}$ を $\tilde{\beta}$ で置き換えたものによる信頼区間は同じオーダーの被覆確率を持つ．

実際に信頼区間を構成するときは $\hat{b}^* = \sum_i (p_i^* - p_i^0)\hat{f}_i(p^0)$ の近似を次のように計算する．$-n^{-1} \leq \delta \leq n^{-1}$ に対して

$$p_j^{i_1 i_2, \delta} = \begin{cases} n^{-1} + \delta & if \quad j = i_1 \\ n^{-1} - \delta & if \quad j = i_2 \\ n^{-1} & if \quad j \neq i_1 \text{ or } i_2 \end{cases}$$

とおき，任意に固定された k に対して $\hat{b}_\delta^* = \delta^{-1} \sum_{i \neq k} (p_i^* - p_i^0)\{\hat{f}(p^{ik,\delta}) - \hat{f}(p^0)\}$ とすると

$$\hat{b}_\delta^* \to \hat{b}^* \quad \text{as} \quad \delta \to 0.$$

$T_b^*, p_{ib}^* (1 \leq b \leq B)$ を b 番目の再抽出標本に基づく T^*, p_i^* に対応するものとすると

$$\hat{Q}_{\delta,B}(\beta) = B^{-1} \sum_{b=1}^{B} I\left[T_b^* \leq \hat{x}(\beta) + \delta^{-1} \sum_{i \neq k} (p_{ib}^* - p_i^0)\{\phi_B^{ik,\delta} - \hat{f}(p^0)\}\right]$$

が $\hat{Q}(\beta)$ の近似となる．ここで $\phi_B^{ik,\delta}$ は $\hat{f}(p^{ik,\delta})$ の不偏推定量である．これを使って信頼区間が構成できる．

シュミレーションの結果でも，この方法はダブル・ブートストラップ法と同じぐらいの精度があることが分かった．ブートストラップ法は適用範囲の広い手法で，推測の精度も高いことから様々な問題に応用され，理論的な性質も明らかにされている．その意味でも，この方法は今後広く応用されることが期待される．

ニューロ判別モデルとリサンプリング法

大阪電気通信大学 総合情報学部 辻谷 将明
バイエル薬品株式会社 開発部門 越水 孝

1．はじめに

本報告では、階層型ニューラルネットワークを判別問題(Discriminant Problem)に適用する。これはロジスティック判別分析の拡張と考えられる。データ解析の目的が予測にあるなら、母集団構造の非線形性を摘出するだけでは不十分で、適切な非線形モデルで記述しなければならない。出力値の確率的解釈によってネットワーク尤度を構成し、尤度原理に基づく統計的推測を行う(越水・辻谷,1998:米谷ら,1998)。ブートストラップ法(Tsujitani&Koshimizu,2000)やクロス・バリデーション法などのリサンプリング法を援用し、*i*)隠れユニット数の決定、*ii*)適合度検定、*iii*)誤判別率のバイアス補正および*iv*)モデル診断(影響分析)を試みる。

2. ニューロ判別モデル

多群判別問題の実際例として、地方銀行の格付けデータを取上げる(川野,1999)。具体的には、4種類の説明変数(リスク管理債権比率、有価証券比率、地域的必要性、資本構成)に基づいて4段階に格付け(評価)された119例のデータに、階層型ニューラルネットワークを適用する。図1のような階層型ニューラルネットによって判別問題を取扱う場合、ある入力パターンを提示したときの出力値は、出力層のあるユニットに入るベイズの事後確率と考えられる。I 個の説明変数から成る入力値 $\{x_1, x_2, ..., x_I\}$ が得られたとき、隠れ層の第 j ユニットの活性値

$$u_j = \sum_{i=0}^{I} \alpha_{ij} x_i,\ j = 1, \cdots, H; x_0 \equiv 1 \tag{1}$$

が定まる。(1)式の活性値 u_j にシグモイド(ロジット)変換

$$f(u_j) = \frac{1}{1 + \exp(-u_j)} \tag{2}$$

を施すと、隠れ層の第 j ユニットの出力値

$$y_j = f(u_j) \tag{3}$$

が決まる。これが、出力層への活性値

$$v_k = \sum_{j=0}^{H} \beta_{jk} y_j; y_0 \equiv 1$$

となる。この v_k にソフトマックス(一般化ロジット)変換を施すと、最終出力

$$o_k = g(v_k) = \frac{\exp(v_k)}{\sum_{k=1}^{K} \exp(v_k)}, k = 1,2,...,K \tag{4}$$

が得られる。

K 個の群について、初期標本 $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K\}$ が得られたとする。ここに、$\boldsymbol{X}_k = \{\boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)}, \cdots, \boldsymbol{X}_{n_k}^{(k)}\}$ の各要素 $\boldsymbol{X}_i^{(k)}$ は、I 個の入力値 $\{x_1, x_2, ..., x_I\}$ で構成されている。$d\left(= 1,2,...,D; D = \sum_{k=1}^{K} n_k\right)$ 番目の教師値を $t_k^{<d>}$、出力値を $o_k^{<d>} = o_k^{<d>}(\boldsymbol{X}; \theta)$、$\theta = \{\alpha, \beta\}$ すると、尤度は

$$L(\boldsymbol{X}|\theta) = \prod_{d=1}^{D} \coprod_{k=1}^{K} \{o_k^{<d>}\}^{t_k^{<d>}},\ \sum_{k=1}^{K} o_k^{<d>} = 1 \tag{5}$$

で与えられる。(5)式を最大にする $\hat{o}_k^{<d>}$ (すなわち、$\hat{\alpha} = \{\hat{\alpha}_{ij}\}, \hat{\beta} = \{\hat{\beta}_{jk}\}$)が最尤推定量(MLE)となる。

3. リサンプリング法

(1)ブートストラップ法

ニューラルネットは、母集団モデルの近似であるため、真のモデルと想定したそれは分離している

(model misspecified)と考えるべきである(Anders&Korn,1999)。この点から、母集団に真のモデルが含まれていることを前提にした、AIC によるモデル選択は妥当ではない。そのため、*TIC*(Shibata,1997)や*NIC*(Murata et al.,1999)が提案されてきた。

競合するモデルが複数個あるとき、対数尤度の比較によって、モデルを選択すると、自由なパラメータ数の大きいモデルほど選ばれやすい。*EIC*(Ishiguroet al.,1997;Konishi&Kitagawa,1991)は、対数尤度のバイアスをブートストラップ法を用いて直接推定している。ブートストラップ法を用いると

$$C^* = E_X\left\{\ln L\left(X^* \mid \hat{\theta}\left(X^*\right)\right) - \ln L\left(X^* \mid \hat{\theta}\left(X^*\right)\right)\right\} \tag{6}$$

によってバイアスの近似値が得られる。ただし、$\boldsymbol{X}^* = \left\{\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_K^*\right\}$はブートストラップ標本である。$\hat{\theta}(\boldsymbol{X}), \hat{\theta}\left(\boldsymbol{X}^*\right)$はそれぞれ初期標本$\boldsymbol{X}$、およびブートストラップ標本$\boldsymbol{X}^*$に基づく$\theta$の推定量とする。このとき、ブートストラップ法に基づく情報量規準

$$EIC = -2\ln L\left(\boldsymbol{X} \mid \hat{\theta}(\boldsymbol{X})\right) + 2C^* \tag{7}$$

が得られ、*EIC* が最小のモデルを最適モデルとして選択することができる。なお、本報告では、分離サンプリングを採用する。

モデルの適合度は、**逸脱度**(deviance)を用いて検証できる。しかし、本報告で取上げているグループ化されていない二値データの場合、(8)式の漸近カイ二乗性は成立たない(Collett,1991;Landwehr et al.,1984)。そこで、ブートストラップ法に基づく棄却点の算出を試みる。

更に、判別分析では初期標本に基づいて何らかの判別ルールを構築するが、この初期標本に対する誤判別率は**見かけ上の誤判別率**と呼ばれる。**実際の誤判別率**は、初期標本に基づいて判別ルールが得られたという条件のもとで将来観測されるデータに対して予測を誤ってしまう確率である。これを見かけ上の誤判別率で置き換えて推定することは、実際の誤判別率を過小に推定する傾向がある。ブートストラップ法を用い、見かけ上の誤判別率のバイアス補正を行う(Gong,1986)。

(2) クロス・バリデーション法

二群判別問題の実際例として、糖尿病データ(Prechett,1994)を取上げる。768 人の糖尿病の疑いがある患者について、５つの検査項目が観察されている。このデータに関し、クロス・バリデーション法を適用する。

初期標本$\boldsymbol{X} = \left\{\boldsymbol{X}^{<1>}, \boldsymbol{X}^{<2>}, \cdots, \boldsymbol{X}^{<D>}\right\}, \boldsymbol{X}^{<d>} = \left\{x_1^{<d>}, x_2^{<d>}, \cdots, x_I^{<d>}; t^{<d>}\right\}$から i 番目の$\boldsymbol{X}^{<i>}$を消去した訓練標本を$\boldsymbol{X}_{[i]} = \left\{\boldsymbol{X}^{<1>}, \boldsymbol{X}^{<2>}, \cdots, \boldsymbol{X}^{<i-1>}, \boldsymbol{X}^{<i+1>}, \cdots, \boldsymbol{X}^{<D>}\right\}$とする。このとき、クロス・バリデーション規準

$$CV規準 = -2\sum_{i=1}^{D} \ln L\left(\boldsymbol{X}^{<i>} \middle| \theta\left(\boldsymbol{X}_{[i]}\right)\right) \tag{8}$$

が最小になる隠れユニット数を最適とする。この規準は、*TIC* と漸近同等である(Shibata,1997)。

クロス・バリデーション法による誤判別率のバイアス補正について述べる。訓練標本$\boldsymbol{X}^{<i>}$に基づく判別ルールを$\eta_{X_{[i]}}$とする。この判別ルールを用いたとき、訓練標本$\boldsymbol{X}^{<i>}$に基づく入力データ$\left(x_1^{<i>}, \cdots, x_I^{<i>}\right)$の予測値を$\eta_{X_{[i]}}\left(x_1^{<i>}, \cdots, x_I^{<i>}\right)$としたとき

$$Q\left(t^{<i>}; \eta_{X_{[i]}}\left(x_1^{<i>}, \cdots, x_I^{<i>}\right)\right) = \begin{cases} 1: \text{誤判別} \\ 0: \text{正判別} \end{cases}$$

と定義すると、クロス・ヴァリデーション法によるバイアス補正された誤判別率は、

$$\frac{1}{D}\sum_{i=1}^{D} Q\left(t^{<i>}; \eta_{X_{[i]}}\left(x_1^{<i>}, \cdots, x_I^{<i>}\right)\right)$$

となる。

また、、一つの観測値を除去することによるﾓﾃﾞﾙ適合度への影響を調べるため逸脱度に関する

$$\Delta Dev_{[d]} = Dev - Dev_{[d]} \tag{9}$$

を用いる。ここに、$Dev = 2\left[L\left(\max|X\right) - L\left(\hat{\theta}|X\right)\right]$, $Dev_{[d]} = 2\left[L\left(\max|X_{[d]}\right) - L\left(\hat{\theta}_{[d]}|X_{[d]}\right)\right]$とする。$d$ 番目の観測値を除去して推定したθの値を$\hat{\theta}_{[d]}$としたとき、$L\left(\hat{\theta}_{[d]} \middle| X_{[d]}\right)$は$X_{[d]}$に対する対数尤度である。$Dev$の$d.f. = n - p$, $Dev_{[d]}$の$d.f. = n - p - 1$より、$\Delta Dev_{[d]}$の$d.f. = n - p - (n - p - 1) = 1$となる。よって、$\Delta Dev_{[d]}$は自由度１のカイ二乗分布に従う。

論理アルゴリズムに基づく条件探索とその改良

Exploration of Conditional Patterns by Logical Algorithm and its Improvement

立教大学・社院　河野康成　立教大学・社　山口和範　創価大学・工　浅野長一郎

1. 序

ネットワーク社会への移行に伴い、大量データの分析も活発化し、データマイニングに関する研究も定着しつつある。マシーンラーニング、パターン認識、統計、データベース、視覚化など学際的な分野からデータマイニング技法が生み出されている[1]。

Quine-McClusky アルゴリズム[3]を用いた方法[2](以下 QM 法)は、大量データの原因条件の分類に極めて有効である。QM 法は、二値型目的変数の応答確率に従って原因条件を分類する手法である。対象データは、基本的に二値型多変量データ行列を用い、Y を目的変数、$X_1, X_2, \ , X_p$ を説明変数、T $(0 \leqq T \leqq 1)$をターゲット値とする。このようなデータを対象として、任意に指定した目的変数の正応答確率に応じて、全体集団の中から条件をみたす部分集団を検出する。分析目的は、目的変数$(Y)=1$ となる確率がターゲット値(T)以上となる説明変数群$(X_1, X_2, \ , X_p)$における条件縮約形の探索である。さらに、標本誤差を考慮し、正応答確率の信頼区間を推定しながら部分集団の検出を縮約式で求める。

方法手順は、以下のような 5 段階で行う。

ステップ 0：ターゲット値の設定
　求めたい部分集団の正応答確率を設定
ステップ 1：カットオフ値の設定
　説明変数が同一内容の出力変数値 1(0)をもつパターンの区切り値を設定
ステップ 2：ターゲット集団の指定
　Quine-McClusky アルゴリズムにより取得された縮約式から求められた部分集団を指定
ステップ 3：信頼区間の作成
　各集団に対する応答確率の信頼区間を作成
ステップ 4：終了判定
　（改善の余地があればステップ 1 に戻る）

まず、ステップ 0 で、目標とするターゲット集団の正応答確率であるターゲット値を設定する。この値は、分析者が目的に従って設定するものであり、一意的なものではない。次に、ステップ 1 で、説明変数が同一内容の入力変数値をもつパターンについて、目的変数である出力変数値 1(0)を再コードするためにカットオフ値を設定する。ここで目的変数 Y の代りに Y を設け、各パターンがカットオフ値以上の場合 $Y=1$、未満の場合 $Y=0$ とする。カットオフ値の初期値は通常ターゲット値と同パーセンテージに設定する。この理由は、各原因条件における正応答確率の最小値が最低限カットオフ値よりも高い数値を取るからである。そして、ステップ 2 で、Quine-McClusky アルゴリズムによって求められた縮約式（簡略化された各項について論理記号 OR を用いて示したもの）に従って、ターゲット集団を指定する。次に、ステップ 3 で、各集団に対する応答確率の信頼区間の作成し、その下限を求める。最後に、ステップ 4 で、ターゲット集団の値が信頼区間の下限より大きいか小さいかをチェックし、改善の余地があれば「ステップ 1」に戻る。

QM 法は、二値型目的変数の応答確率に従って、必要条件・十分条件を探索し、大量デー

タ分析に効果的である。しかし、変数が多くなる場合は、結果となる縮約式が冗長となり、信頼区間を作成する意味もなくなる可能性がある。本研究では、これらの問題を解決するために、変数の次元を制約するアルゴリズムを提唱した。

2. 制約付アルゴリズム

説明変数が多い大量データの場合、前処理段階において、変数削減をするのもひとつの手段である。たとえば、相関の高い変数を選択するという方法も考えられるが、交互作用を見落とす可能性がある。そこで、変数削減とは別の方法として、周辺分割表に QM 法を用いる制約付アルゴリズムを考案した。制約付アルゴリズムの手順は以下の通りとした。

1. 次元数(p)の設定
2. 低次の周辺分割表における全てのパターンに対する QM 法の適用
3. 2 の結果に対して Quine-McClusky アルゴリズムによる縮約

まず、次元数の設定は、分析者の分析目的に従って決定される。次に、n 変数のデータに対して、次元数を p 変数に設定し、${}_nC_p$ 通りの周辺分割表における全てのパターンに対して QM 法が適用される。さらに、それぞれの結果に対して、Quine-McClusky アルゴリズムによる縮約が行なわれるが、既に 2 の段階で信頼区間の下限を満たしているターゲット集団だけが選出されているためこの段階では信頼区間作成の必要はない。

この制約付アルゴリズムについて数値例を示して考察を行った。その結果、n 変数のデータを次元数 p に縮約することが可能となった。最終的に、変数を制約した縮約式は、元の変数のデータを直接 QM 法で分析した縮約式に対応できており、信頼区間の観点から見ても十分であることが判明した。なお、このアルゴリズムは、変数・データ数共に多くなれば多くなるほど効果的であると考えられる。

3. まとめ

変数削減の方法に対して、総当りの QM 法は、全ての交互作用を考慮している点で優れている。また、結果の冗長性については、提唱方法を用いることで改善ができた。しかしながら、現状では、変数が多い場合、この方法をとると計算時間が多大にかかってしまうという欠点を抱えている。

今後は、計算時間や不用な変数を削除することなどを踏まえて経験的に考察していく必要がある。

参考文献

[1] Cavena, P., Hadjinian P., Stadler, R., Verhees, J. & Zanasi, A.: *Discovering Data Mining, From Concept to Implementation*, Prentice-Hall, Inc. NJ (1997)

[2]河野康成: Quine-McCluskyアルゴリズムを用いた目的別対象集団の探索手法. 応用社会学研究 43, 95-101 (2001)

[3] Quine, W. V.: The Problem of Simplifying Truth Functions, *American Mathematics Monthly, October, 1952, 59*, 521-531 (1952)

Mining Frequent Patterns from Graph Structured Data

ISIR., Osaka University, Hiroshi Motoda

Data having graph structure are abound in many practical fields such as molecular structures of chemical compounds, information flow patterns in the internet, DNA sequences and its 3D structures, and inference patterns (program traces of reasoning process). Thus, knowledge discovery from structured data is one of the major research topics in recent data mining and machine learning study. In this work we focus on mining typical patterns in a graph structure data. By "typical" we mean frequently appearing subgraphs in the whole graph data. Conventional empirical inductive learning methods and association rules in data mining use an attribute-value table as a data representation language, and cannot treat a graph structure directly. What makes the problem difficult is that subgraph isomorphism is known to be NP complete.

We have taken two different approaches. One is a quite simple heuristic based approach (hereafter called GBI, Graph-Based Induction). It is based on the notion of pairwise chunking and no backtracking is made (thus approximate). Its time complexity is almost linear with the size of graph but as is evident it only gives approximate solutions. The other is an approach based on Apriori's bottom up algorithm[Agrawal94] which is extended to handle graph structure (hereafter called AGM, Apriori-based Graph Mining). Its time complexity is exponential to the size of graph but it gives exact solutions. Both methods can handle directed/undirected, colored/uncolored graphs with/without self loop and with colored/uncolored links. The exact method can further handle subgraph patterns partitioned into multiple parts.

GBI The central intuition behind is that a pattern that appears frequently enough is worth paying attention to and may represent an important concept (which is implicitly embedded in the input graph). In order to extract frequently appearing subgraphs, stepwise pair expansion (repeated pairwise chunking) is performed by repeating the following three steps[Matsuda01]: 1) If there are patterns identical to the chunked pattern in the graph, rewrite each of them to a single node of a new label. 2) Extract all linked pairs in the input graph. 3) Select the most frequent pair and register it as the pattern to chunk. Each time we perform the pairwise chunking, we keep track of link information between nodes in order to be able to restore the original graph(s) or represent the extracted patterns in terms of the original nodes. This is realized for a directed graph by keeping two kinds of node information: "child node information" (which node in the pattern the link goes to) and "parent node information" (which node in the pattern the link comes from). For an undirected graph, it is first converted to a directed graph by imposing a certain fixed order to node labels. The time complexity of the algorithm is $O(CP + NL)$ where N, L, P, C respectively denote the total number of nodes in the graph, the average number of links going out of one node, the number of different kinds of pairs in the graph, and the number of different kinds of chunked patterns derived from the graph data. Experimentally this is shown to be almost linear with the size of graph.

AGM Unlike GBI, AGM aims to perform a complete search admitting its exponential time complexity. We follow the idea used in Apriori algorithm that is based on the monotonicity of support: all the subset of a frequent itemset must be frequent itemsets. Extending this principle to subgraph extraction allows us to enumerate all the frequent subgraphs for small graphs. The graph structured data is transformed into an adjacency matrix, and a candidate subgraph of size k is constituted by two frequent subgraphs of size $k-1$ that share a size $k-2$ subgraph. Notion of normal form is introduced not to generate redundant candidates. A recursive algorithm is developed to identify a canonical form for isomorphic normal forms. Using this algorithm in a bottom up manner, association rules having a support and a confidence greater than user specified thresholds can be enumerated where condition and conclusion are both subgraphs. One important strength of this approach is that a subgraph is not necessarily a connected graph. As expected, the computation time is only linear to the number of transactions (input graphs) but exponential to the size of each transaction and a support threshold.

Extracting Patterns from Chemical Compound Data To see how both GBI and AGM work, two domains of chemical compound was chosen. One is to predict chemical carcinogenicity of organic

chlorides, and the other is to predict mutagenicity of aromatic or heteroaromatic nitro compounds. Some examples of extracted patterns (rules) for the first task are shown in Fig. 1. Both AGM and GBI worked as expected. AGM can indeed extract disconnected subgraphs. When the minimum support level is lowered, computational complexity of AGM become a problem, but no problem arises for GBI. The number of extracted patterns is less for GBI but for this type of problem GBI seems to work well and can be used to extract at least important patterns.

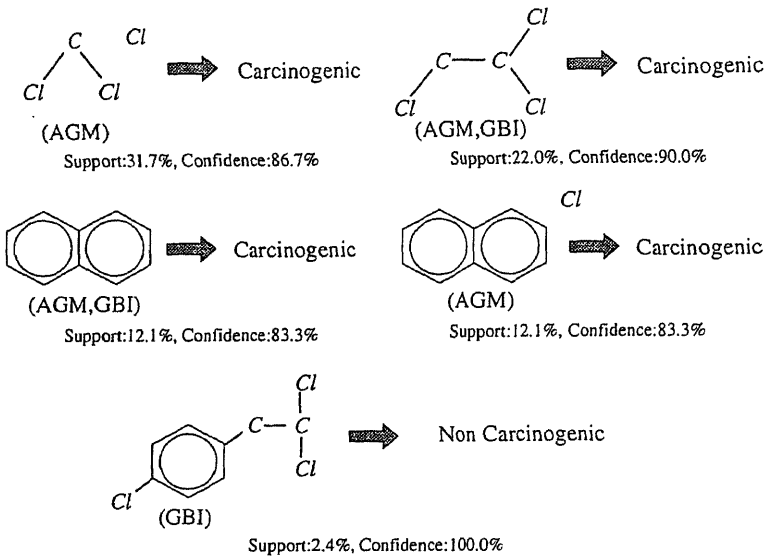


Figure 1: Example rules derived from chemical carcinogenesis data

Summary We showed two methods that can extract typical patterns from graph structured data. One is based on repeated pairwise chunking which is very fast, obtains approximate solutions, and runs almost linearly to the size of graph. The other is based on the extended Apriori algorithm which is complete in search but slow, and runs exponentially to the size of graph. Both are complementary to each other in many respects. Initial results of their applications to chemical compound analyses suggest that they are useful in identifying important substructure that are responsible to carcinogenicity and mutagenicity.

References

[Agrawal94] R. Agrwal and R. Srikant. First Algorithms for Mining Association Rules. *Proc. of the 20th VLDB Conference*, pp. 487–499, 1994.

[Inokuchi0001] A. Inokuchi, T. Washio, H. Motoda, K. Kumazawa and N. Arai. Fast and Complete Mining Method for Frequent Graph Patterns (in Japanese). *Journal of Japanese Society for Artificial Intelligence*, Vol. 15, No. 6, pp. 1052–1063, 2000.

[Matsuda01] T. Matsuda, H. Motoda and T. Washio. Graph-Based Induction for General Graph Structured Data and Its Applications (in Japanese). *Journal of Japanese Society for Artificial Intelligence*, Vol. 16, No. 4, pp. 363–374, 2001.