2003年日本オペレーションズ・リサーチ学会
春季研究発表会

$m \times n$ 分割表の近似数え上げスキームの提案

東京大学　*来嶋秀治　KIJIMA Shuji
01605000　東京大学　松井知己　MATSUI Tomomi

1. はじめに

2元分割表は正の整数からなる行和と列和を持ち、表中の値として非負整数をとる表（行列）である。2元分割表は医療統計の分野などで統計データを扱うのに用いられる。与えられた行和および列和を満たす2元分割表の個数を厳密に数える問題は、行数が2の時でさえ#P完全であることが知られている。

2000年にDyer and GreenhillによってMCMC（マルコフ連鎖モンテカルロ）法を用いた$2 \times n$分割表の個数を求める近似解法が提案された。この方法は非常に直感的だが、$m \times n$分割表に適用した場合、精度や偏りの理論的考察が困難となる。本報告では、分割表の個数に関する性質を考慮することでDyer and Greenhillの方法を改良する。さらに我々の方法は$m \times n$分割表に拡張することが可能である。本報告では、我々の手法によって得られる推定量の精度とその期待値の真の値からの偏りの大きさについて議論する。

2. 近似解法

整数（非負整数、正整数）全体の集合をそれぞれ$\mathbb{Z}$（$\mathbb{Z}_+$、$\mathbb{Z}_{++}$）で表すことにする。ベクトル$r = (r_1, \ldots, r_m) \in \mathbb{Z}_+^m$と$s = (s_1, \ldots, s_n) \in \mathbb{Z}_+^m$は正整数$N \in \mathbb{Z}_{++}$に対して、$\sum_{i=1}^m r_i = \sum_{j=1}^n s_j = N$を満たすとする。行和$r$および列和$s$をもち、非負整数を表値にとる$m$行$n$列の2元分割表全体の集合$\Sigma_{r,s}$を、

$$\Sigma_{r,s} \stackrel{\text{def.}}{=} \left\{ X \in \mathbb{Z}_+^{m \times n} \;\middle|\; \sum_{j=1}^{n} X_{ij} = r_i \quad (1 \le \forall i \le m), \quad \sum_{i=1}^{m} X_{ij} = s_j \quad (1 \le \forall j \le n) \right\}$$

で定義する。但し、X_{ij}は分割表Xのi行j列の値を表す。明らかに分割表の個数$|\Sigma_{r,s}|$は行と列を置き換えても変わらない。したがって、一般性を失うことなく$m \le n$とする。部分集合$\Omega_i \subset \Sigma_{r,s}$を$\Omega_i = \{X \in \Sigma_{r,s} | X_{in} \ge \lceil s_n/m \rceil\}$で定義する。また、添え字$k$は$r_k = \max\{r_1, \ldots, r_m\}$を満たすとする。この時、

$$\tilde{r} \stackrel{\text{def.}}{=} (r_1, \ldots, r_k - \lceil s_n/m \rceil, \ldots, r_m),$$

$$\tilde{s} \stackrel{\text{def.}}{=} \begin{cases} (s_1, \ldots, s_{n-1}, \lfloor \frac{m-1}{m} s_n \rfloor), & s_n > 1, \\ (s_1, \ldots, s_{n-1}), & s_n = 1, \end{cases}$$

を定義する。明らかに$|\Omega_k| = |\Sigma_{\tilde{r},\tilde{s}}|$が成り立つ。もし、$\rho = |\Omega_k|/|\Sigma_{r,s}|$と$|\Sigma_{\tilde{r},\tilde{s}}|$がわかれば$|\Sigma_{r,s}| = |\Sigma_{\tilde{r},\tilde{s}}|/\rho$を計算できる。しかし、一般に$\rho$および$|\Sigma_{\tilde{r},\tilde{s}}|$を求めることは困難である。いま、$\Sigma_{r,s}$上で一様標本抽出が可能ならば、$\rho$の値の推定にモンテカルロ法を適用できる。すなわち$M$回の標本抽出を行い、$U$個の分割表が$\Omega_k$に含まれるならば、$(U+1)/(M+1)$を$\rho$の推定量とする。こうして、もとの問題に比べてサイズの小さな分割表$\Sigma_{\tilde{r},\tilde{s}}$の数え上げ問題に帰着させることが出来る。この手続きを繰り返して、分割表のサイズが2行2列になるまで小さくする。2×2分割表の個数は定数時間で求めることができるので、再帰的に$|\Sigma_{r,s}|$を求めることができる。

3. 分割表の個数に関する諸定理

我々のスキームでは、問題のサイズを小さくする目的で部分集合Ω_kを定義した。この定義には二つの重要な意味がある。一つは行の添え字kの決め方で、もう一つはサイズ縮小の値$\lceil s_n/m \rceil$である。これは二つの背反な要求から来る。まず、多項式時間のアルゴリズムを得るために、効率的に問題のサイズ縮小を行う必要がある。また、比の値ρが小さすぎるとρの推定量が誤差に敏感になるのでρはある程度十分大きくなければならない。この目的で我々は次の定理を示した。

定理 1 添え字$k \in \{1, \ldots, m\}$は$r_k = \max\{r_1, \ldots, r_m\}$を満たすとする。この時、$|\Omega_k| \ge \frac{1}{m}|\Sigma_{r,s}|$が成り立つ。

この定理は次の補題を用いて示すことができる。

補題 2 $\boldsymbol{r}, \boldsymbol{r}' \in \mathbb{Z}_+^m$ および $\boldsymbol{s} \in \mathbb{Z}_+^n$ は $|r_1 - r_2| \leq |r_1' - r_2'|$, $r_i = r_i' (i = 3, \ldots, m)$, $\sum_{i=1}^m r_i = \sum_{i=1}^m r_i' = \sum_{j=1}^n s_j = N$, を満たすとする。この時、二つの集合 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ と $\Sigma_{\boldsymbol{r}',\boldsymbol{s}}$ の間で $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \geq |\Sigma_{\boldsymbol{r}',\boldsymbol{s}}|$ が成り立つ。

定理 1 から、我々のスキームに現れる理論比 ρ は $\rho \geq 1/m$ を満たすことが分かる。

4. 推定量の誤差と偏り

ここでは、我々の近似解法で得られる近似解の誤差と偏りについて議論する。提案した近似解法中で $m \times n$ 分割表の一様標本抽出を仮定した。しかし、一般に $m \times n$ 分割表の一様標本抽出は困難であるため、我々の提案するスキームでは近似一様標本抽出を仮定する。集合 Ω 上の分布 π と ν のあいだの総分布距離 $\mathrm{d_{TV}}(\pi, \nu)$ を、$\mathrm{d_{TV}}(\pi, \nu) = \frac{1}{2} \sum_{x \in \Omega} |\pi(x) - \nu(x)|$ で定義する。いま、集合 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ 上の一様分布を π で表す。任意の正数 $\varepsilon < 1$ に対して、ある近似的な一様標本抽出法が存在して、標本は $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ 上の分布 ν に従って抽出されるものとし、分布 ν は総分布距離 $\mathrm{d_{TV}}(\pi, \nu) \leq \varepsilon/(6mR)$ を満たすとする。但し、R は Z を得るために必要な問題のサイズ縮小の回数とする。この標本抽出法を用いて標本数 $M = 108mR^2\varepsilon^{-2} \ln(2R/\delta)$ のモンテカルロ法を行い、前節で提案した手法における比 ρ の推定量 $\hat{\rho}$ を計算する。逐次この推定比を用いて $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ の推定量 Z を求める。すなわち我々のスキームは多項式回の問題縮小と多項式個の標本抽出で終了する。この時、推定量 Z に関して次の二つの定理が成り立つ。

定理 3 推定量 Z は

$$\Pr\left[(1-\varepsilon)|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \leq Z \leq (1+\varepsilon)|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|\right] \geq 1 - \delta$$

を満たす。

定理 4 推定量 Z は

$$\frac{|\mathrm{E}[Z] - \Sigma_{\boldsymbol{r},\boldsymbol{s}}|}{|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|} \leq \frac{\varepsilon}{4} + \mathrm{e}^{-90R^3\varepsilon^{-2}\ln(2R/\delta)} < \left(\frac{1}{4} + \frac{1}{10^{27}}\right)\varepsilon$$

を満たす。

5. 結論と課題

我々のスキームは、$m \times n$ 分割表数え上げ問題に対して多項式回の問題縮小と多項式個の標本抽出で誤差の大きさと偏りの幅を確率論的に押さえられた近似解を与える。2002 年 Cryan et al. が行数が定数の 2 元分割表に対する heat bath マルコフ連鎖が rapid mixing であることを示した。もちろんこれを我々のスキームに適用すると、我々のスキームも行数固定の $m \times n$ 分割表数え上げ問題に対する多項式時間確率論的近似解法となる。しかし、一般の $m \times n$ 分割表に対する多項式時間近似一様標本抽出法の存在性は現時点では未解決である。

参考文献

[1] M. Cryan, M. Dyer, L. A. Goldberg, M. Jerrum, and R. Martin, "Rapidly mixing Markov chains for sampling contingency tables with constant number of rows," *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science (FOCS)*, (2002), pp. 711–720.

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