数理解析研究所講究録
第1849巻 2013年 105-108

複数ストリーム間の特徴比較に対する乱択アルゴリズム

A randomized algorithm for comparison between streams

園田 尚人 Naoto Sonoda
山内 由紀子 Yukiko Yamauchi
来嶋 秀治 Shuji Kijima
山下 雅史 Masafumi Yamashita

九州大学
Kyushu University

1 はじめに

スーパーのPOSシステムやパケットモニタリング，センサーデータ等，大規模なストリームデータを出力する計算機システムの重要性は，情報社会の発展に伴って増大してきている．このようなシステムが出力する複数のデータが与えられたとき，その特徴を分析することは基本的な問題の一つである．このようなシステムにおいては，一般的に非常に大きなデータが出力されるため，全てのデータを記憶して分析を行うようなアルゴリズムは空間複雑性の観点から困難である．この点に留意して，未知のサイズに対して正常に動作するストリーミングアルゴリズムの設計が必要となる．

ストリーミングアルゴリズムについては，データの種類数のカウントでさえ厳密解を計算できない場合があることが知られている [4]．そこで，近似アルゴリズムを設計し理論的保障を与えることが必要となる．データの特徴量のひとつである頻出アイテムの検知については，緒方ら [3] による結果が存在する．緒方らはサンプリングを用いることで $O(\log\log N)$bit の記憶領域のみを使用して動作する頻出アイテム検知アルゴリズムを設計し，その保証を与えた．

本研究では，複数のストリームが与えられたとき，各ストリームで特徴的に出現するアイテムについて知ることを考える．例えば，複数のアクセスログがある場合に，あるログでは大量に出現しているが別のログでは出現が少ないようなものを発見したい．そこで，ストリーム間のアイテムの分布の差を調査する省スペースなアルゴリズムを提案する．提案手法の基本的なアイデアは Cormode らの提案したリゾーバサンプリング [1] であり，このサンプルに対して分布の差の推定に対する理論的保障を与えた．さらに，アルゴリズムが $O(\log\log N)$ の記憶領域で動作することに言及した．

本論文では，まず2章で問題の定義について述べる．次に3章では，リゾーバサンプリングについて記述する．その後，4章で分布の差の推定法と解析について述べる．最後に5章で結論と今後の課題を記述する．

2 問題の定義

本章では，本研究で扱う問題について定義する．Σ をアイテムの全集合とする．簡単のために Σ を有限集合とし，各要素は $\sigma \equiv \log|\Sigma|$ ビットで識別できるとする．Σ の要素から成るアイテム列 $\boldsymbol{x} = x_1, ..., x_N$ をストリームと呼ぶ．$\boldsymbol{x}$ に対して，$f(s;\boldsymbol{x}) \in \{0, ..., N\}$ を $\boldsymbol{x}$ 中のアイテム $s \in \Sigma$ の出現回数とする．また，それぞれ長さが N_A，N_B のストリーム $\boldsymbol{x}_A$，$\boldsymbol{x}_B$ が与えられたとき，あるアイテム s についての分布の差を $f_A(s) = f(s;\boldsymbol{x}_A)$，$f_B(s) = f(s;\boldsymbol{x}_B)$ として以下のように定義する．

$$\left|\frac{f_A(s)}{N_A} - \frac{f_b(s)}{N_B}\right|$$

本研究では，少ない記憶領域でを用いて上記の分布の差を近似するアルゴリズムについて議論する．長さ N_A，N_B は非常に大きいと仮定する．すなわち，全てのアイテムを記憶して厳密な分布の差を求めることは困難である．そこで，各ストリーム中より c 個のアイテムを一様乱択した集合から分布の差を推測することを考える．

3 リゾーバサンプリング

本節では準備として，Cormode ら [1] のリゾーバサンプリングアルゴリズムについて述べる．いま，$c \in \mathbb{Z}_{>0}$ をアルゴリズムに対して与えられた定数とし，K，K' を高々c 個のアイテムからなる多重集合とする．また，二つの多重集合 A と B にたいして，$A \uplus B$ を A と B の和とする．このときアルゴリズムは以下のように与えられる．

Algorithm 1

1: **Set** $K, K' := \emptyset$. **Set** " exponent " $h := 0$.
2: **Read** an input x_i if it exists,
otherwise **goto** 6.
3: **Add** x_i in K or K' with probability 2^{-h}.
Add x_i to K or K' with the same probability
4: **If** $|K| = c$, then
increment h by one.
Set $K' := \emptyset$.
for each x_i in K,
move x_i from K to K' with probability $\frac{1}{2}$.
5: **Goto** 1.
6: **Output** c items uniformly random
from $K \uplus K'$.

Algorithm1 は以下の定理 3.1 を満たす．

定理 3.1

Algorithm1 にて，各アイテム x_i $(i \in \{1, ..., n\})$ は等確率で出力される．

証明

まず，アイテムが出力される確率について述べる．アルゴリズム終了時，各アイテムがそれぞれ 2^{-h} の確率で K または K' に入っていることを示すのは容易である．$K \uplus K'$ から c 個のアイテムを一様乱択するのであるから，各アイテムが出力される確率は等確率になっているといえる． ■

したがって，Algorithm1 は $n \geq c$ の場合，c 個の一様サンプルを出力する．アルゴリズム中で n の値を必要としないため，サイズが未知のストリームに対しても動作することができる．以下の章では，本サンプリングアルゴリズムで得られたサンプルに対して議論をおこなう．

4 分布の差の推定

前章で述べたアルゴリズムにより，我々はストリーム中からの c 個の一様サンプルを手に入れることができた．本章ではこのサンプルよりストリーム中のアイテムの分布の差を近似する手法について提案する．まず，このサンプルは以下の観察が成り立つことを述べておく．

観察 4.1

長さ N のストリーム中から c 個のアイテムを一様乱択したとき，あるアイテム s が乱択される個数 k は超幾何分布にしたがう．すなわち以下の式が成り立つ．

$$\Pr[k = i] = \binom{f(s;\boldsymbol{x})}{i}\binom{N - f(s;\boldsymbol{x})}{c-i}\Big/\binom{N}{c}$$

2 つのストリーム $\boldsymbol{x}_A$，$\boldsymbol{x}_B$ 中からそれぞれサンプルした c 個のアイテム中であるアイテム s の出現数を $k_A(s)$，$k_B(s)$ とする．$k_A(s)$，$k_B(s)$ はそれぞれ観察 4.1 より超幾何分布にしたがう確率変数である．このサンプルの $\left|\frac{f_A(s)}{N_A} - \frac{f_b(s)}{N_B}\right|$ に対する推定量として

$$\left|\frac{k_A(s)}{c} - \frac{k_B(s)}{c}\right|$$

を用いることにする．また，表記の簡略化のために以下では，$X = \frac{f_A(s)}{N_A} - \frac{f_b(s)}{N_B}$，$Y = \frac{k_A(s)}{c} - \frac{k_B(s)}{c}$ を用いることとする．

式 $|Y|$ が $|X|$ を確率的に近似することを示したい．しかしながら，差が小さい場合にはこの近似比率はサンプリングによる誤差に非常に敏感である．例えば，$|X| = 0$ のときは，$k_A(s) = k_B(s)$ となれば分布の差を近似できているが，一つでもサンプル数が異なった場合には近似比が無限大となってしまう．この点に留意して以下の定理が導かれる．

定理 4.2

あるアイテム s について，$\max\left\{\frac{f_A(s)}{N_A}, \frac{f_B(s)}{N_B}\right\} = \theta$，$\min\left\{\frac{f_A(s)}{N_A}, \frac{f_B(s)}{N_B}\right\} = \alpha\theta$ とする．ただし θ，α はそれぞれ $(0,1)$ の値とする．任意のパラメータ $\epsilon \in (0,1)$，$\delta \in (0,1)$ に対し，サンプルサイズ c が

$$c \geq \frac{(1+\alpha) - \theta(1+\alpha^2)}{\epsilon^2\theta(1-\alpha)^2\delta} \tag{1}$$

を満たすとき，確率 $1-\delta$ 以上で

$$1 - \epsilon \leq \frac{|Y|}{|X|} \leq 1 + \epsilon \tag{2}$$

となる．

本定理は，ある程度以上頻出であるアイテムで，2つのストリーム中での分布の差がある程度大きい場合には，サンプリングによって得られる分布の差の近似について精度保証が得られることを意味する．すなわち，一方のストリームでは多数出現するが他方では出現しづらいような，ストリーム間で特徴的であるアイテムについて分布の差が推定できることを述べている．

以下では定理 4.2 を示す．

証明

一般性を失うことなく $f_A(s) \geq f_B(s)$ を仮定する．すなわち，$\theta = \frac{f_A(s)}{N_A}$，$\alpha\theta = \frac{f_B(s)}{N_B}$ とする．まず $\mathrm{E}[Y/|X|]$ に着目する．k_A，k_B は超幾何分布にしたがうことから以下の式を導くことができる．ただし，$|X| = \theta(1-\alpha)$ であることに注意する．

$$\begin{aligned}\mathrm{E}[Y/|X|] &= \frac{1}{c\theta(1-\alpha)}\mathrm{E}[k_A - k_B] \\ &= \frac{1}{c\theta(1-\alpha)}\mathrm{E}[k_A] - \mathrm{E}[k_B]) \\ &= \frac{1}{c\theta(1-\alpha)}(c\theta - c\theta\alpha) = 1\end{aligned}$$

この事実とチェビシェフの不等式より以下を得る．

$$\begin{aligned}&\Pr[1-\epsilon \leq |Y|/|X| \leq 1+\epsilon] \\ \geq\ &\Pr[1-\epsilon \leq Y/|X| \leq 1+\epsilon] \\ =\ &\Pr[|Y/|X| - 1| \leq \epsilon] \\ \geq\ &1 - \frac{\mathrm{Var}[Y/|X|]}{\epsilon^2}\end{aligned}$$

したがって，$\mathrm{Var}[Y/|X|]$ さえわかれば式 (2) の成り立つ確率を述べることができる．k_A と k_B は独立な確率変数であるから，$\mathrm{Var}[Y/|X|]$ について次の式が成り立つ．

$$\begin{aligned}&\mathrm{Var}[Y/|X|] \\ =\ &\frac{1}{(c\theta(1-\alpha))^2}(\mathrm{Var}[k_A] + \mathrm{Var}[k_B]) \\ \leq\ &\frac{1}{(c\theta(1-\alpha))^2}(c\theta(1-\theta) + c\alpha\theta(1-\alpha\theta)) \\ =\ &\frac{(1+\alpha) - \theta(1+\alpha^2)}{c\theta(1-\alpha)^2}\end{aligned}$$

よって，この式と c の仮定より，

$$\Pr[1-\epsilon \leq |Y|/|X| \leq 1+\epsilon] \geq 1-\delta \qquad (3)$$

が成り立つ．したがって題意を満たす． ■

いま，あるしきい値 θ，α を与える．また θ'，α' を

$$\begin{aligned}\theta' &:= \max\left(f_A(s)/N_A, f_B(s)/N_B\right) \\ \alpha' &:= \min\left(f_A(s)/N_A, f_B(s)/N_B\right)/\theta' \qquad (\theta' > 0)\end{aligned}$$

と定義する．このときに $\theta' \geq \theta$ かつ $\alpha' \leq \alpha$ である全てのアイテム s について分布の差を推定したいと考えよう．すなわち，しきい値より頻出であり，かつ分布の差の大きなアイテム全てについての推定値を求めるのである．実は，このような s については c を式 (1) のようにおいたときに式 (3) を必ず満たす．なぜならば，$\mathrm{Var}[Y/|X|]$ が θ について単調減少かつ α について単調増加だからである．さらに，このような s の個数は高々2θ 個しかないことを用いると次の定理を導くことができる．

定理 4.3

θ，$\alpha \in (0,1)$ を与えられた定数とする．任意のパラメータ $\epsilon \in (0,1)$，$\delta \in (0,1)$ に対し，サンプルサイズ c が

$$c \geq \frac{2\left((1+\alpha) - \theta(1+\alpha^2)\right)}{\epsilon^2\theta^2(1-\alpha)^2\delta}$$

を満たすとき，確率 $1-\delta$ 以上で $\theta' \geq \theta$ かつ $\alpha' \leq \alpha$ である全てのアイテム s について式 *(2)* を満たす．

最後に，本手法において使用した記憶領域について言及する．

定理 4.4

$O(\log\log n)bit$ の記憶領域のみを用いて，分布の差を推定することが高確率で可能である．

証明

本手法において使用する記憶領域は Algorithm1 で用いる領域のみである．Algorithm1 では記憶領域として K，K' と h を使用している．K を記憶するには $c \cdot \sigma$bit あれば十分である．K' は最悪の場合には非常に大きな記憶領域を必要とする場合もある．しかし，Chernoff 上界 [2] を用いた解析により，$4c \cdot \sigma$bit 以上必要となる確率は，c について指数的に小さくなる．また，h を記憶するには $\log h$bit 必要である．ここでまた Chernoff 上界を用いた解析を行うと，h は非常に高い確率で $\log n$ に近似することがわかる．以上から必要な記憶領域は高確率で $O(c) + O(\log h) = O(\log\log n)$ となる． ■

5 まとめと今後の課題

本研究では，大規模なデータストリームから限られた記憶領域のみを用いてアイテムの分布を比較する手法を提案した．そのさい，アイテムが頻出であることを表すしきい値θと分布の差のθに対する割合を表すしきい値αに着目した．その結果，$O(\log\log n)$の記憶領域のみを用いて，$\theta' \geq \theta$かつ$\alpha' \leq \alpha$である全てのアイテムについて，確率的に分布の差を近似することに成功した．

しかしながら，本研究では$\theta' \geq \theta$かつ$\alpha' \leq \alpha$を満たさないアイテムについてはなにも言及できていない．そのため，本手法に該当するアイテムについては確かに推定できるのであるが，該当しないアイテムの挙動が分からないために，どの推定値か正しいものか判断することが困難であるという問題点が存在する．この点の解決が今後の課題である．

解決手段の一つは，$\theta' \geq \theta$を満たすアイテムを頻出アイテム検知アルゴリズムを用いて取り出す方法である．これは緒方ら [3] のアルゴリズムと組み合わせることが可能であると予想している．

参考文献

[1] G. Cormode, S. Muthukrishnan, K. Yi, and Q. Zhang, Optimal sampling from distributed streams, PODS '10, 77–86

[2] M. Mitzenmacher and E. Upfal, Probability and Computing: Randomaized Algorithms and Probabilitistic Analysis, Cambridge University Press, 2005.

[3] M. Ogata, Y. Yamauchi, S. Kijima, and M. Yamashita, A randomized algorithm for finding frequent elements in streams using O(loglogN) space, Lecture Notes in Computer Science 7074 (ISAAC2011), 514–523.

[4] 徳山 豪, オンラインアルゴリズムとストリームアルゴリズム, 共立出版, 2007.