LOS ALAMOS
NATIONAL LABORATORY

Concentration of the Hypergeometric Distribution

Don Hush and Clint Scovel
Modeling, Algorithms and Informatics Group, CCS-3
Mail Stop B265
Los Alamos National Laboratory
Los Alamos, NM 87545
(dhush,jcs)@lanl.gov

LANL Technical Report: LA–UR–03–1353

Report Date: March 10, 2003

Abstract

In this paper we provide an improved concentration of measure theorem for the hypergeometric distribution.

Concentration of the hypergeometric distribution $K(n_1, n, m)$, counting the number of defectives obtained when n_1 items are selected randomly without replacement from n of which m are defective, is important in learning theory. For example Vapnik's proof of Theorem 4.1 in (Vapnik, 1998), controlling the difference between generalization error and training error uniformly over some hypothesis class, utilizes the concentration of the hypergeometric distribution $K(n_1, 2n_1, m)$. His proof can easily be modified to utilize the concentration of the more general $K(n_1, n, m)$. In addition, Cannon (Cannon, Ettinger, Hush, & Scovel, 2002a) use the concentration of $K(n_1, n, m)$ to obtain bounds on estimation error in a learning problem where the freedom to specify n might be useful. Serfling (Serfling, 1974) provides bounds on the concentration of the hypergeometric distribution. However, extending Vapnik's proof on page 163 in (Vapnik, 1998) of the concentration of $K(n_1, n, m)$ for $n = 2n_1$ to general n provides a substantial improvement over Serfling's result. This improvement is similar to that obtainable for binomial sampling when the binomial probability p is small, (e.g. see (McDiarmid, 1998)). In this paper we state and prove this new concentration theorem and compare with Serfling's result.

We now state and prove our main theorem.

Theorem 0.1. *Let $K = K(n_1, n, m)$ denote the hypergeometric random variable describing the process of counting how many defectives are selected when n_1 items are randomly selected without replacement from a population of n items of which m are defective. Let $\gamma \geq 2$.*

Then

$$\mathcal{P}(K - E(K) > \gamma) < e^{-2\alpha_{n_1,n,m}(\gamma^2-1)}$$

and

$$\mathcal{P}(K - E(K) < -\gamma) < e^{-2\alpha_{n_1,n,m}(\gamma^2-1)}$$

where $\alpha_{n_1,n,m} = \max\left(\left(\frac{1}{n_1+1} + \frac{1}{n-n_1+1}\right), \left(\frac{1}{m+1} + \frac{1}{n-m+1}\right)\right)$.

Before we proceed with the proof, let us compare the first bound with Serfling's result (Serfling, 1974)

$$\mathcal{P}(K - E(K) \geq \gamma) \leq e^{-\frac{2n}{n_1(n-n_1+1)}\gamma^2}. \tag{1}$$

Comparison with the second bound is the same. We compare the rate coefficients $\alpha_s = \frac{n}{n_1(n-n_1+1)}$ from the Serfling's bound with $\alpha_{n_1,n,m}$. If we denote $\acute{\alpha} = \frac{1}{n_1+1} + \frac{1}{n-n_1+1}$ then it is clear that $\alpha_{n_1,n,m} \geq \acute{\alpha}$. It is not hard to show that for all $n \geq 1$, $n_1 \geq 1$ and $n_1 \leq n$ we have

$$\frac{\alpha_s}{2} < \acute{\alpha} < 2\alpha_s$$

so that the exponential rate of Theorem 0.1 is no worse than half that Serfling's bound (1). It is clear that for large n_1 and n these coefficents are very close. In addition, when $2n_1 \geq n$, we obtain $\acute{\alpha} \geq \alpha_s$. On the other hand when $m < \min(n_1, n - n_1)$ or $m > \max(n_1, n - n_1)$

$$\alpha_{n_1,n,m} = \frac{1}{m+1} + \frac{1}{n-m+1} \tag{2}$$

which can be $\mathcal{O}(n_1)$ times as large as $\frac{n}{n_1(n-n_1+1)}$ resulting in a much stronger concentration. For example when $m = 0$ and $n = 2n_1$, $\alpha_{n_1,n,m} = \frac{1}{m+1} + \frac{1}{n+1-m}$ is greater than $\frac{n_1}{2}$ times as large as $\alpha_s = \frac{n}{n_1(n-n_1+1)}$. We now proceed with the proof of the Theorem.

Proof. Let $C_n^m = \binom{n}{m}$ and let $n_2 = n - n_1$. We follow the proof in Vapnik (Vapnik, 1998), page 163, which is valid when $n_1 = n_2$. The hypergeometric distribution has the probability

$$p(k) = \frac{C_m^k C_{n-m}^{n_1-k}}{C_n^{n_1}}$$

for $\max(0, m-n_2) \le k \le \min(n_1, m)$. Let $S = \max(0, m-n_2)$ and $T = \min(n_1, m)$.

Consider

$$q(k) = \frac{p(k+1)}{p(k)} = \frac{m-k}{k+1} \cdot \frac{n_1 - k}{n_2 + k + 1 - m}.$$

It is clear that q monotonically decreases as k increases. Define $d(k) = \sum_{i=k}^{T} p(i)$. Then

$$d(k+1) = \sum_{i=k+1}^{T} p(i) = \sum_{i=k}^{T-1} p(i+1) = \sum_{i=k}^{T-1} q(i)p(i)$$

which is less than

$$q(k)\sum_{i=k}^{T-1} p(i) \le q(k)\sum_{i=k}^{T} p(i) = q(k)d(k)$$

by the monotonicity of q. Therefore,

$$d(k+1) < q(k)d(k).$$

Consequently, for any $S \le j < k \le T-1$,

$$d(k) < d(j)\prod_{i=j}^{k-1} q(i)$$

and since $d(j) \le 1$, we obtain

$$d(k) < \prod_{i=j}^{k-1} q(i)$$

Consider k a continuous variable for the moment. $\kappa = \frac{m(n_1+1)-(n_2+1)}{2+n}$ is the unique point such that $q(\kappa) = 1$. Consider the function $Q(t) = q(\kappa + t)$. Then

$$\ln Q(0) = 0$$

and

$$\frac{d}{dt}(\ln Q(t)) = -\frac{1}{m-\kappa-t} - \frac{1}{\kappa+t+1} - \frac{1}{n_1-\kappa-t} - \frac{1}{n_2+\kappa+t+1-m}.$$

Since

$$\frac{1}{\tau}\frac{1}{a-\tau} \geq \frac{4}{a}$$

when $a > 0$ and $a \geq \tau \geq 0$, if we consider pairing the first with the fourth term and the second with the third term, we get the bound

$$\frac{d}{dt}(\ln Q(t)) \leq -4\Big(\frac{1}{n_1+1} + \frac{1}{n_2+1}\Big).$$

Likewise, we can pair the first with the second and the third with the fourth to obtain the bound

$$\frac{d}{dt}(\ln Q(t)) \leq -4\Big(\frac{1}{m+1} + \frac{1}{n-m+1}\Big).$$

Consequently $\frac{d}{dt}(\ln Q(t)) \leq -4\alpha_{n_1,n,m}$. Dropping subscripts and writing $\alpha = \alpha_{n_1,n,m}$ we have

$$\frac{d}{dt}(\ln Q(t)) \leq -4\alpha.$$

Integration from s up to t where $t \geq s$ and $S \leq \kappa + s \leq T$ and $S \leq \kappa + t \leq T$ yields

$$\ln Q(t) - \ln Q(s) \leq -4\alpha(t-s). \tag{3}$$

In particular $\ln Q(t) \leq -4\alpha t$ for $t \geq 0$ and so

$$\ln d(k) < \sum_{i=j}^{k-1} \ln q(i) = \sum_{i=j}^{k-1} \ln Q(i-\kappa) \leq -4\alpha \sum_{i=j}^{k-1}(i-\kappa)$$

when $\kappa \leq j \leq k-1$, but since we can bound the sum like so

$$\sum_{i=j}^{k-1}(i-\kappa) \geq \int_{i=j}^{k}(i-\kappa-1)di = \frac{1}{2}(k-\kappa-1)^2 - \frac{1}{2}(j-\kappa-1)^2$$

we have the bound

$$\ln d(k) < -4\alpha\Big(\frac{1}{2}(k-\kappa-1)^2 - \frac{1}{2}(j-\kappa-1)^2\Big).$$

Now choose $j = \lceil\kappa\rceil$ the smallest integer greater than or equal to κ. Then $(j-\kappa-1)^2 = (\lceil\kappa\rceil-\kappa-1)^2 \leq 1$ so that we have

$$\ln d(k) < -2\alpha\Big((k-\kappa-1)^2 - 1\Big).$$

Denote $E = E(K) = \frac{mn_1}{n}$. The constraint $k \geq j+1$ is guaranteed when $k \geq E+2$ since $E \geq \kappa$ implies that $E+1 \geq \lceil\kappa\rceil$. Let

$$\Delta = E - \kappa = \frac{1+n_2}{2+n} + \frac{m(n_1-n_2)}{n(2+n)}$$

and note that we have the bounds

$$\frac{1}{2+n} \leq \frac{\min(n_1, n_2) + 1}{2+n} \leq \Delta \leq \frac{\max(n_1, n_2) + 1}{2+n} < 1.$$

Substitute $\kappa = E - \Delta$ to obtain

$$\ln d(k) < -2\alpha\Big((k - E + \Delta - 1)^2 - 1\Big)$$

but since $\Delta < 1$ we obtain

$$\ln d(k) < -2\alpha\Big((k - E)^2 - 1\Big).$$

Consequently, we have proven that

$$\mathcal{P}(K - E(K) > \gamma) < e^{-2\alpha(\gamma^2 - 1)}.$$

To get the other side of the bound we proceed in a similar fashion. Now instead define $d(k) = \sum_{i=S}^{k} p(i)$. Then

$$d(k-1) = \sum_{i=S}^{k-1} p(i) = \sum_{i=S+1}^{k} p(i-1) = \sum_{i=S+1}^{k} \frac{1}{q(i-1)} p(i)$$

which is less than

$$\frac{1}{q(k-1)} \sum_{i=S+1}^{k} p(i) \leq \frac{1}{q(k-1)} \sum_{i=S}^{k} p(i) = \frac{1}{q(k-1)} d(k)$$

by the monotonicity of q. Therefore,

$$d(k-1) < \frac{1}{q(k-1)} d(k).$$

Consequently, for any $S \leq k \leq j \leq T - 1$,

$$d(k) < d(j) \prod_{i=k}^{j} \frac{1}{q(i)}$$

and since $d(j) \leq 1$, we obtain

$$d(k) < \prod_{i=k}^{j} \frac{1}{q(i)}$$

Now we set $t = 0$ in inequality 3 to obtain

$$-\ln Q(s) \leq -4\alpha(0 - s)$$

for $s \leq 0$. Consequently,

$$-\ln q(i) \leq -4\alpha(\kappa - i)$$

for $i \leq \kappa$. Then we obtain

$$\ln d(k) < -\sum_{i=k}^{j} \ln q(i) \leq -4\alpha \sum_{i=k}^{j} (\kappa - i)$$

for $S \leq k \leq j \leq \kappa$, but since we can bound the sum

$$\sum_{i=k}^{j} (i-\kappa) \leq \int_{i=k}^{j+1} (i-\kappa) di = \frac{1}{2}(j+1-\kappa)^2 - \frac{1}{2}(k-\kappa)^2$$

we have the bound

$$\ln d(k) < -4\alpha\Big(\frac{1}{2}(k-\kappa)^2 - \frac{1}{2}(j+1-\kappa)^2\Big).$$

We now choose $j = \lfloor \kappa \rfloor$, but since $(\lfloor \kappa \rfloor + 1 - \kappa)^2 \leq 1$ we obtain

$$\ln d(k) < -2\alpha\Big((k-\kappa)^2 - 1\Big).$$

The constraint $k \leq j$ is satisfied when $k \leq E - 2$ since $\kappa \geq E - 1$ implies that $\lfloor \kappa \rfloor - 2$. Since $\Delta \geq 0$ it follows that $k - E \leq k - \kappa$ and we can write

$$\ln d(k) < -2\alpha\Big((k-E)^2 - 1\Big).$$

Therefore we have shown that

$$\mathcal{P}(K - E(K) < -\gamma) < e^{-2\alpha(\gamma^2 - 1)}$$

and the proof is finished.

♦

References

Cannon, A., Ettinger, M., Hush, D., & Scovel, C. (2002a). Machine learning with data dependent hypothesis classes. *Journal of Machine Learning Research*, *2*, 335–358.

McDiarmid, C. (1998). Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, & B. Reed (Eds.), *Probabilistic methods for algorithmic discrete mathematics* (pp. 195–248). Springer-Verlag.

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, *2*(1), 39–48.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: John Wiley and Sons, Inc.