Lecture - 35

Thursday, 3 November 2016 (5:10 - 6:00)

Species Accumulation (Discovery) Problem

1 Species Accumulation Curve

The species accumulation (discovery) curve is a plot recording the cumulative number of species of living things recorded in a particular environment as a function of the cumulative effort expended searching for them.

2 Chao 92 Estimator

This section presents a method by Chao et al. [1] to estimate the number of unique classes (elements) in a population where the elements have unknown probabilities. The method is based on the concept of sample coverage. There have been some previous works trying to answer this problem, however, they either work in case of equal probability or assume some distribution of the elements. The current work rests on top of the previous work and provides a decent estimation in the general case, i.e, without assuming any particular underlying distribution.

Terminologies Following are the terminologies used in the estimation technique:
N = The number of unique species (classes)
(Assume that the classes are indexed by $1, 2, 3, \ldots, N$.)
n = The size of the sample
p_i = The probability that an element belongs to ith class.
X_i = Total number of elements of the ith class observed in the sample, where $i = 1, 2, \ldots, N$.
f_i = The number of classes having i elements
f_i can be represented in terms of X_i as:

$$f_i = \sum_{j=1}^{N} I[X_j = i]$$

where I is the indicator random function.
D = The number of observed classes. Obviously $D \leq N$
D can be represented in terms of X_i as:

$$D = \sum_{i=1}^{N} I[X_i >= 1]$$

Or, D can be represented in terms of f_i as:

$$D = \sum_{i=1}^{N} f_i$$

Sample size n can be represented in terms of f_i as:

$$n = \sum_{i=1}^{N} i f_i$$

C = Sample Coverage, which is defined as sum of the probabilities of the observed classes. Alternatively, it is the probability that if you pick a class, and it has been observed already.

Aim: Our aim is to estimate N based on f_i.

Sample Coverage:

1. **Equiprobable Case:** In the equiprobable case, all the N classes have equal probabilities, i.e,

   $$p_1 = p_2 = \cdots = p_N = 1/N$$

   In this case, sample coverage is given as,

   $$C = \frac{D}{N}$$

   Hence, N can be estimated as:

   $$N = \frac{D}{C}$$

2. **General Case:** Since sample coverage is the sum of the probabilities of the observed classes, it is given as:

   $$C = \sum_{i=1}^{N} p_i I[X_i > 0]$$

   where I is the indicator random function telling whether a particular class has been observed or not. Note that in an unknown area, we obviously do not know the probabilities of the classes. Therefore, we may use the Good-Turing Estimator [2] for estimating the sample coverage, which is given as:

   $$\hat{C} = 1 - f_1/n \tag{1}$$

where $\hat{C}$ is the estimator of C.

Estimating N for the equiprobable case:
Since we know that in this case, $C = D/N$, we can estimate N as $\hat{N}$ as follows:

$$N = \frac{D}{\hat{C}}$$

where $\hat{C}$ can be estimated using the Good-Turing estimator as defined earlier and D can be easily checked in the given sample.

Chao92 Estimator (General Case):

Theorem 1. *Assume that a random sample of size n is drawn from a population of N elemets with fixed elements probabilities $p = (p_1, p_2, \ldots, p_N)$, $\sum p_i = 1$. Let $p_1, p_2, \ldots, p_N$ have a mean $\bar{p} = \sum_i p_i/N = 1$ and coefficient of variation (CV) $\lambda = [\sum_i (p_i - \bar{p})^2/N]^{1/2}/\bar{p}$. ($\lambda = 0$ when all p_i's are equal. Then, the formula for estimating the number of unique classes in a population with unknown probabilities of classes is given by*

$$\boxed{N = \frac{D}{C} + \frac{n(1-C)}{C}\lambda^2} \tag{2}$$

where λ is given by,

$$\max\left\{\frac{\frac{D}{C}\sum i(i-1)f_i}{[n(n-1)]} - 1, 0\right\} \tag{3}$$

Proof. In order to estimate N, let us first calculate the expected values of D and C, i.e. $E(D)$ and $E(C)$ respectively.
E(D): Comparing this to balls and bins problem, finding expected number of observed classes = expected no. of bins that have atleast one ball, where n is the number of balls and N is the number of bins. = N - no. of bins having no ball.
p_i is the probability of a ball going to ith bin.
$(1 - p_i)$ is the probability of a ball not going to ith bin.
$\sum_1^n (1 - p_i)^n$= No. of bins having no balls.
Hence, number of bins having atleast one ball, and hence $E(D)$ may be written as:

$$E(D) = N - \sum_1^n (1 - p_i)^n$$

Similarly, we can estimate the expected value of C. *Computing the expected value of C, i.e., $E(C)$ was given as a tutorial question.*
The term $E(D)/E(C)$, after rearrangement and using certain estimators for λ, i.e. Coefficient of variation provides the formula for N.

3 An Application of Chao92 Estimator

We further discussed a work that applies Chao 92 estimator in a real world scenario [3]. This work uses the concept of Species Accumulation in order to find the completeness of a crowdsourced enumeration query.

In paid crowdsourced environments like *Amazon Mechanical Turk (AMT)*, it becomes important for the administrators/requesters to know: "When is the query result set complete?", i.e. when should we stop taking results from the users. This is because there is time and cost associated with waiting, which needs to be minimized. Therefore, we see a clear tradeoff between the completeness of the query and the time/cost.

Since the formula for estimating the number of species in a population (Chao 92 estimator) is already known, the authors make use of the same formula. However, when used on Crowdsourced enumeration queries, Chao92 shows results that are overestimated. The authors state reasons for this deviation in terms of the differences between the characteristics of species accumulation and a crowdsourced environment. They then make necesary changes into Chao92 estimator to enable it to be used in such environments.

Differences between Species Accumulation and Crowdsourced Environments

1. Species Accumulation samples are based on *with* replacement from an unknown distribution, whereas in crowdsourced enumerations, users provide answers from an underlying distribution *without* replacement.
2. *Worker Skew/Presence of Streakers:* Workers who do significantly more work are called *streakers*. Their presence leads to *worker skew (WS)*. WS causes the arrival rate of new answers to be more rapid that leads the estimator to overestimate.

Improved Estimator (For Crowdsourced Enumeration) As we see that Chao92 estimator makes use of $\hat{C}$ which is estimated in terms of f_1, which is the number of classes having only one element, therefore it is heavily influenced by the presence of rare items in the sample. However, we know that in a crowdsourced environment, when these rare items appear too quickly, which happens quite oftenly in a crowdsourced environment, this leads to overestimation of the total number of unique answers. The authors, therefore, modify this estimator caused by overabundance of f_1 answers, which basically come due to the presence of streakers. These streakers do not give the opportunity to provide answers to other workers which may increase f_2 and f_3 etc. instead of f_1. The modified estimator is designed to identify such workers (called outliers) which lead to a sharp increase in f_1. This is done by altering the f-statistic. We define outliers as the workers that lie outside two standard deviations away from the mean of all workers (W). The modification basically involves changing the formula for f_1, which is done as follows:

1. First the mean and standard deviation for each worker is calculated by excluding the worker under consideration and considering the rest of $W - 1$ workers, as follows:

$$\hat{x_i} = \sum_{\forall j, j \neq i} \frac{f_1(j)}{W - 1}$$

$$\hat{\sigma}_i = \sqrt{\sum_{\forall j, j \neq i} \frac{(f_1(j) - \bar{x_i})^2}{W - 1}}$$

2. $\tilde{f_1}$ is either kept the original f_1 or if it is outside two standard deviations from the mean (i.e. it is the case of a streaker), it is then modified and taken as $2\hat{\sigma}_i + \bar{x_i}$, i.e.,

$$\tilde{f_1}(i) = \sum_i min(f_1(i), 2\hat{\sigma}_i + \bar{x_i})$$

3. The modified estimator is same as the equation 2 except that in the formula for λ^2 (equation 3), f_i is replaced by $\tilde{f_i}$. For example, in the simplest case where $\lambda^2 = 0$, $\hat{N}_{crowd}$ becomes,

$$\hat{N}_{crowd} = \frac{D}{\hat{C}} \tag{4}$$

Now, we know that $\hat{C} = 1 - \frac{f_1}{n}$, which implies, $\hat{C} = \frac{n - f_1}{n}$.
Putting the value of $\hat{C}$ and then $\tilde{f_1}$ in equation 3, we get,

$$\boxed{\hat{N}_{crowd} = \frac{Dn}{n - \sum_i min(f_1(i), 2\hat{\delta} + \bar{x_i})}}$$

Bibliography

[1] A. Chao and S.-M. Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.

[2] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[3] B. Trushkowsky, T. Kraska, M. J. Franklin, and P. Sarkar. Crowdsourced enumeration queries. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 673–684. IEEE, 2013.