Meta-analysis and big data in the case of binary data

Stephan Morgenthaler*
École polytechnique fédérale de Lausanne, Lausanne, Switzerland - stephan.morgenthaler@epfl.ch

Abstract

The term big data often refers to the analysis of data collected from disparate sources, some informative (but of varying quality), others heavily biased or even totally uninformative. One of the methods advocated for such data sets is to analyze smallish subsets of the data either randomly or systematically selected and then to combine the partial results into a global answer. This approach is superficially similar to a meta-analysis. In this paper we analyse the feasibility of this approach when considering a binary response variable. We assume that the observation is either of high quality or of low quality. Low quality means that the observations have undergone a biasing mechanism. We show that even if we know the type of each observation, we cannot do much better with the combined data than with the high quality data alone.

Keywords: Meta-analysis, combining data sets, subset analysis, bias.

1. Simple Subset Analysis

One possible strategy for analysing large databases is to select manageable subsets, analyse them separately and combine the results to reach an overall conclusion. This is similar to the data combination that underlies a meta-analysis of a set of clinical trials. As an example, consider the elementary case of binary observations $x_i \in \{0, 1\}$ (success/failure) and the assessment of the success probability $p = P(\text{success})$.
If a trial with n_k subjects results in x_k successes, the success probability is usually estimated as

$$\hat{p}_k = \frac{x_k}{n_k} \text{ or as } \hat{p}_k = \frac{x_k + 0.5}{n_k + 1}\,.$$

To combine K such estimators, one may use the weighted mean

$$\hat{p}_{\text{meta}} = \frac{\sum_{k=1}^{K} \hat{p}_k/\hat{v}_k}{\sum_{k=1}^{K} 1/\hat{v}_k}\,,$$

where $\hat{v}_k = \hat{p}_k(1 - \hat{p}_k)/n_k$ is the estimated variance of the estimated success probability.
If we assume independence of the trials and treat $\hat{v}_k$ as a constant, the variance of the weighted mean is approximated by

$$\text{Var}(\hat{p}_{\text{meta}}) = \frac{\sum_{k=1}^{K} \text{Var}(\hat{p}_k)/\hat{v}_k^2}{\left(\sum_{k=1}^{K} 1/\hat{v}_k\right)^2} = \frac{1}{\sum_{k=1}^{K} n_k/(\hat{p}_k(1 - \hat{p}_k))}\,,$$

which is of order $O(N^{-1})$, with N denoting the total sample size.
When using a broad mixture of data sources this approach poses serious problems, foremost in the form of biases. Unless care is taken, it is not evident what $\hat{p}_{\text{meta}}$ estimates. This danger holds for any meta-analysis, where one has to question the utility of mixing carefully executed clinical trials with trials of dubious quality and origin. It is equally true in big data applications.

2. A model for binary data and its likelihood

Consider a case, where we combine n independent observations of $X \sim \text{Bernoulli}(p)$ with N independent observations of $Y \sim \text{Bernoulli}(P)$ with

$$P = p(1 - \pi) + (1 - p)\pi = (1 - 2\pi)p + \pi = p + \pi(1 - 2p)\,.$$

Here π is a (small) probability of misclassifying a success as a failure and vice versa. The random variable X represents the high quality observations, whereas Y provides merely a biased assessment of p. The interest

centers on estimating p and the question we want to answer is whether a large sample N of cheap observations can help us in estimating p.
If we use the combined sample in a straightforward manner, we obtain the estimate $(x+y)/(N+n)$, where x and y denote the number of successes in the first sample and second sample, respectively. This estimates the probability in the mixed population $p+(P-p)(1-n/(n+N))$. If we only use the high quality data, then x/n is an unbiased estimate of p.
The log likelihood, under independence of the two sources, is

$$x\log(p)+(n-x)\log(1-p)+y\log(P)+(N-y)\log(1-P)\,.$$

In the parametrization (p,π) of this problem, the likelihood equations are

$$\begin{aligned}\frac{x}{\hat{p}} &= \frac{n-x}{1-\hat{p}}\\ \frac{y}{\hat{p}(1-2\hat{\pi})+\hat{\pi}}(1-2\hat{\pi}) &= \frac{N-y}{1-\hat{p}(1-2\hat{\pi})-\hat{\pi}}(1-2\hat{\pi})\end{aligned}$$

whose solution is

$$\hat{p}=\frac{x}{n}\,.$$

This means that absent further information about π, even a huge second part of the data cannot help in estimating p.
If we knew something about π, for example, if we are willing to assume that it is no more than 0.1. Then we could estimate $\hat{p}$ in an ad-hoc manner from $\hat{P}=y/N$ as

$$\tilde{p}=\frac{\hat{P}-0.1}{1-2\times 0.1}\,,$$

which would exploit the large amount of additional data. The likelihood equation for known π is

$$\frac{x}{\hat{p}}-\frac{n-x}{1-\hat{p}}+\frac{y(1-2\pi)}{\hat{p}(1-2\pi)+\pi}-\frac{(N-y)(1-2\pi)}{1-\hat{p}(1-2\pi)-\pi}=0\,,$$

which gives a polynom of degree three in $\hat{p}$. The root of interest lies between $\hat{p}=x/n$ and $\tilde{p}$.
If all the observations are simply mixed together before analysis, we do not know who is who and the likelihood becomes

$$\sum_{i=1}^{N+n}\log\left(\frac{n}{N}p^{x_i}(1-p)^{1-x_i}+\left(1-\frac{n}{N}\right)P^{x_i}(1-P)^{1-x_i}\right)\,,$$

where $x_1,\ldots,x_{n+N}$ is the sequence of binary observations. This function is maximized by all pairs $(\hat{p},\hat{\pi})$ that are feasible in the sense that

$$(1-2\hat{\pi})\hat{p}+\hat{\pi}=y/N\,.$$

3. Conclusions

Augmenting binary data with biased observations is not of much help. Even if the biasing mechanism is known up to a single parameter. Only if we can guess the value of this parameter or learn about it from other sources, is it possible to construct estimators with much reduced mean-squared-error.
Even biases of order $O(1/n)$ – more commonly encountered in statistics – can cause problems in subset analyses, because in the combination of K estimates based on subsets of size n, they become large when compared to the variance of order $O(1/(Kn))$ and can lead to invalid conclusions.