Performance Metric Proofs for fabl/vabl

Brian Kundinger

April 4, 2024

1 INITIAL THOUGHTS

It seems to me that base Fellegi Sunter is bound to have a higher false positive rate as the size of the problem grows.

- Assuming there are at maximum K records in A that match with record $j \in B$. For one to one matching, this means $K = 1$.
- There are at least $n_A - K$ nonmatching record pairs for each $j \in B$
- If we're making $n_A n_B$ independent classification decisions it just seems intuitive that we would get more false positives as $(n_A - K)n_B$ grows.
- It also seems intuitive that we would tend to see less false positives when we can only make at most n_B classification decisions.

A possible approach: In FS a record pair gets classified as a match if it has an appropriately high w_{ij}. But in fabl, you need to have the high w_{ij}, and no other $w_{i'j}$ can be greater. This might be tangible place to start?

Does that make sense? Does it seem possible to prove rigorously? Have you thought about this before?

If we could do this, it would give a nice theoretical justification for using the fabl framework over base FS (or practically, using vabl over fastLink).

2 AN ATTEMPT AT RIGOR

Essentially, we would try to prove

$$E[\#\text{ False Positives}|\text{FS}] \geq E[\#\text{ False Positives}|\text{fabl}].$$

We would assume all m and u parameters are the same for each model. We assume there are at most $K = 1$ matches in X_1 for each record in X_2. We can use the correspondence that λ under base Fellegi Sunter is equal to $\frac{\pi}{n_A}$ under fabl.

Left hand side:

$$
\begin{aligned}
E[\#\text{ False Positives}|\text{FS}] &= n_A n_B \times p(\text{False Positive}|\text{FS}) \\
&= n_A n_B \times p\left(\frac{u_{ij}(1-\lambda)}{m_{ij}\lambda + u_{ij}(1-\lambda)} \geq 0.5\right)
\end{aligned}
$$