Bayesian Estimation of Bipartite Matchings for Record Linkage[1]

Mauricio Sadinle
Duke University, Durham, North Carolina, USA - msadinle@stat.duke.edu

Abstract

In this article we are concerned with the most traditional scenario of record linkage, which consists of linking two disparate datafiles containing overlapping information on a set of entities, and it is assumed that each entity is recorded maximum once in each datafile. This is an important task with a wide variety of applications, given that it needs to be solved whenever we have to combine information from different sources. Most statistical techniques currently in use are derived from a seminal paper by Fellegi and Sunter (1969) who formalized procedures that had been used earlier by other researchers. These techniques usually assume independence in the matching status of record pairs to derive estimation procedures and optimal point estimators (e.g. Fellegi-Sunter decision rule). We argue that this independence assumption is unreasonable and target instead a bipartite matching between the two sets of records coming from the two files as our parameter of interest. The Bayesian implementation presented here allows us to incorporate prior information on the quality of the fields in the datafiles, which in turn helps to obtain better results when the datafiles do not share a large amount of identifying information. We demonstrate the improvements of our approach over traditional methodologies in a number of realistic simulation studies.

Keywords: bipartite matching; data integration; data matching; Fellegi-Sunter.

[1]Supported by NSF grants SES-11-30706 to Carnegie Mellon University and SES-11-31897 to Duke University.