Chapter 19 [1]

STATISTICAL RULE INDUCTION IN THE PRESENCE OF PRIOR INFORMATION: THE BAYESIAN RECORD LINKAGE PROBLEM

Dean H. Judson[2]
U.S. Census Bureau
Washington, D.C. 20233
U.S.A.
Email: *Dean.H.Judson@census.gov*

Abstract: This chapter applies the theory of Bayesian logistic regression to the problem of inducing a classification rule. The chapter first describes the classification rule as a decision to link or not to link two records in different databases in the absence of a common identifier. When a training data set of classified cases is available, developing a rule is easy; this chapter expands the application of the technique to situations where a training data set of classified cases is *not* available. The steps are conceptually simple:first fit a logistic regression with latent dependent variable using Bayesian methods, then use the parameter estimates from the best fitting model to derive the equivalent record linkage rule. This chapter first describes the application area of record linkage, followed by a description of the Fellegi-Sunter model of record linkage. The chapter then shows how to estimate the appropriate Bayesian generalized linear model with latent classes, and, using the posterior kernels, determine the final decision rule.

Key Words: Bayesian record linkage, concept learning, latent class analysis, unsupervised learning, Fellegi-Sunter model.

---

[1] Triantaphyllou, E. and G. Felici (Eds.), **Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques**, Massive Computing Series, Springer, Heidelberg, Germany, pp. 655-694, 2006.

[2] This chapter reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This chapter is released to inform interested parties and encourage discussion.