Chapter 16[1]

Predictive Regression Modeling for Small Enterprise Datasets with Bootstrap, Clustering and Bagging

C. Jack Feng and Krishna Erla
Department of Industrial and Manufacturing Engineering
Bradley University, Peoria, Illinois 61625, U.S.A.
Email: cfeng@bradley.edu

Abstract: Most enterprise datasets are large, but some are very small for predictive purposes due to expensive experiments, reduced budget or tight schedule required to generate them. The bootstrap approach is a method used frequently for small datasets in data mining. Numerous theoretical studies have been done on bootstrap in the past two decades but few have applied it to solve real world manufacturing problems. Bootstrap methods provide an attractive option when model selection becomes complex due to small sample sizes and unknown distributions. In principle, bootstrap methods are more widely applicable than the jackknife method, and also more dependable. In this chapter we focus on selecting the best model based on prediction errors computed using the revised bootstrap method, known as the *0.632 bootstrap*. Models developed and selected are then clustered and the best cluster of models is next bagged to provide the minimum prediction errors. Numerical examples based on a small enterprise dataset illustrate how to use this procedure in selecting, validating, clustering, and bagging predictive regression models when sample sizes are small compared to the number of parameters in the model.

***Key Words*:** Bootstrap sampling; Predictive regression; Subset selection; Bagging; Cluster analysis.

---

[1] Liao, T.W. and E. Triantaphyllou, (Eds.), **Recent Advances in Data Mining of Enterprise Data**, *World Scientific*, Singapore, pp. 747-774, 2007.