# Developing Classifiers from Unbalanced Datasets

Konstantinos Falangis and John Glen

Management School, University of Edinburgh

**Abstract**

Credit scoring and related datasets often consist of very unbalanced populations in each class, with the minority class representing a small proportion of the total population. For example, in credit scoring bad cases typically comprise less than 10% of the population, while in fraud detection the proportion of fraud cases may be much smaller than the proportion of non-fraud cases since the probability of fraud is very low (e.g. less than 0.2%). Datasets with extremely unbalanced class sizes create difficulties in developing classifiers because the classification rule may be dominated by the majority class. In this paper we examine methods for dealing with unbalanced class sizes in using statistical and mathematical programming techniques to develop classification models. The application of these methods is illustrated using a number of credit scoring datasets and a fraud detection dataset.