Generalization Method when Manipulating Relational Databases

V. Cariou*,** L. Billard***

*LUNAM University, Oniris, Sensometrics and Chemometrics Laboratory,
Nantes Cedex F-44322 France
veronique.cariou@oniris-nantes.fr
**INRA, Nantes, F-44307, France
***Department of Statistics, University of Georgia, Athens, GA 30602, USA
lynne@stat.uga.edu

Abstract. Contemporary computers generate massive datasets. One way to handle these data is to aggregate them into smaller datasets (with the aggregation criteria dictated by meaningful scientific questions of interest). This paper focuses on aggregations that produce interval datasets. Algorithms are introduced both to build intervals which are typically homogeneous, and to test that such homogeneity pertains. They also test whether or not observations across the resulting intervals are mixtures of uniform distributions rather than the desired single distribution. These include consideration of outlier observations. The methods are illustrated on two datasets.

1 Introduction

Contemporary datasets can be enormous, too large for standard analytic methods to be used directly on the very same computers generating the datasets themselves. Thus, some form of data manipulation is needed in order to transform the original dataset into one that is more manageable for appropriate analyses.

There are many approaches that have been proposed to address this issue. Most have different strengths, most are more applicable to some settings and data types than others; all are useful. Data mining as a broadly based methodology identifies patterns in the dataset, and then delves more deeply into the part of the data responsible for those patterns, perhaps as one operation or perhaps by pattern type. See, e.g., Hand et al. (2001). Another approach is to take a sample of the dataset. One such technique is data squashing whereby the original data are sorted into clusters of like characteristics, with a “representative" pseudo-sample drawn from each cluster. The analysis is conducted on this scaled down sampled dataset. See, e.g., DuMouchel et al. (1999).

A third broad approach developed in the literature deals with aggregating data points, where the criteria for any particular aggregation vary depending on the nature of the scientific questions being asked. The resultant dataset then consists of lists, intervals, histograms, and the like, and fall under the general heading of symbolic data. See, e.g., Bock and Diday (2000)