IDEA GROUP PUBLISHING
701 E Chocolate Avenue, Hershey PA 17033, USA
Tel: 717/533-8845; Fax: 717/533-8661; URL: www.idea-group.com
ITB7257

Chapter VIII

How Size Matters: The Role of Sampling in Data Mining

Paul D. Scott
University of Essex, UK

This chapter addresses the question of how to decide how large a sample is necessary in order to apply a particular data mining procedure to a given data set. A brief review of the main results of basic sampling theory is followed by a detailed consideration and comparison of the impact of simple random sample size on two well-known data mining procedures: naïve Bayes classifiers and decision tree induction. It is shown that both the learning procedure and the data set have a major impact on the size of sample required but that the size of the data set itself has little effect. The next section introduces a more sophisticated form of sampling, disproportionate stratification, and shows how it may be used to make much more effective use of limited processing resources. This section also includes a discussion of dynamic and static sampling. An examination of the impact of target function complexity concludes that neither target function complexity nor size of the attribute tuple space need be considered explicitly in determining sample size. The chapter concludes with a summary of the major results, a consideration of their relevance for small data sets and some brief remarks on the role of sampling for other data mining procedures.

INTRODUCTION

When data mining emerged as a distinct field, it was plausibly claimed that the total quantity of information stored in databases doubled every 20 months (Frawley, Piatetsky-Shapiro & Matheus, 1991). The credo of the new discipline was that the

This chapter appears in the book, *Heuristic and Optimization for Knowledge Discovery*, edited by Ruhul Sarker, Hussein Abbass and Charles Newton.

effort expended on accumulating and storing this prodigious quantity of data should be regarded as an investment that had created a resource ripe for exploitation. Machine learning had produced a number of well-proven techniques for automatically discovering regularities and patterns in data sets. The idea of applying these techniques to find the untapped seams of useful information in these vast deposits of data was the starting point of the new discipline. In the subsequent decade, size appears to have undergone a seismic shift in status: very large databases are now regarded as problematic because it may not be possible to process them efficiently using standard machine learning procedures. The problem is particularly acute when the data is too large to fit into main memory.

There are three basic approaches to dealing with this problem: first, develop new algorithms with more modest space/time requirements; second, use existing algorithms but implement them on parallel hardware (see Freitas & Lavington, 1998 for review); and third, apply the learning procedures to an appropriate sample drawn from the data set.

Machine learning practitioners appear uncomfortable with the idea of sampling; for them, it is what John and Langley (1996) describe as "a scary prospect". Why this should be is something of a puzzle, since sampling theory is a long-established area of study. Standard introductory texts on statistics (e.g. Wonnacott & Wonnacott, 1990) typically include a treatment of those basic aspects of the subject that bear directly on hypothesis testing; Mitchell (1997) covers similar material in a machine learning context. Sampling itself is usually treated in separate texts: Kalton (1983) provides a concise introduction, while Kish (1965) provides a more comprehensive and mathematically grounded coverage.

In this chapter I shall be concerned with one central question: how do you decide how large a sample you need in order to apply a particular data mining procedure to a given data set. In the next section I discuss why sampling is unavoidable and review the main results of basic sampling theory. The following section comprises a detailed examination of the impact of sample size on two well-known data mining procedures: naïve Bayes classifiers and decision tree induction. The next section introduces disproportionate stratification and shows how it may be used to make much more effective use of limited processing resources. This section also includes a discussion of dynamic and static sampling. This is followed by a section devoted to the impact of target function complexity on sample size. The final section provides a summary of the major results, a consideration of their relevance for small data sets and some brief remarks on the role of sampling for other data mining procedures.

SAMPLING

Many people reject the idea of sampling on intuitive grounds. Unfortunately, human intuition is often poor on matters concerning sampling; the erroneous belief that a larger population implies the need for a correspondingly larger sample is very