Representative training sets for classification and the variability of empirical distributions

Saaid Baraty, Dan Simovici

University of Massachusetts Boston
sbaraty@cs.umb.edu, dsim@cs.umb.edu

Abstract. We propose a novel approach for the estimation of the size of training sets that are needed for constructing valid models in machine learning and data mining. We aim to provide a good representation of the underlying population without making any distributional assumptions.

Our technique is based on the computation of the standard deviation of the χ^2-statistics of a series of samples. When successive statistics are relatively close, we assume that the samples produced represent adequately the true underlying distribution of the population, and the models learned from these samples will behave almost as well as models learned on the entire population.

We validate our results by experiments involving classifiers of various levels of complexity and learning capabilities.

1 Introduction

Estimating a sample size that allows the inference of a good model is an important part of the learning process. We seek to determine the minimum size of a sample which is very likely to be a "fair" representative of the underlying population. Models learned from these samples will behave almost as well as models learned on the entire population and any increase in the size of the sample would result in insignificant increases in the quality of the models.

Our goal is to determine sample sizes that are sufficient to ensure that these samples adequately represent the underlying population. These samples are used as training sets for constructing models comparable in performance with those inferred from the entire population, but are cheaper to build.

Sections 2 and 3 describe in detail our approach for finding the size of a sample from a data set and a population respectively. The experimental work is presented in Section 4.

2 Estimating the size of a sample from data

Let $U = \{u_1, u_2, \ldots, u_n\}$ be a set of attributes. The set of possible states for attribute u_i, $\mathrm{Dom}(u_i)$, is assumed to be finite and is commonly referred to as *domain* of u_i. The notion of domain of attributes is extended to sets of attributes by defining the set $\mathrm{Dom}(V)$ for the set of attributes $V \subseteq U$ as $\mathrm{Dom}(V) = \prod_{v \in V} \mathrm{Dom}(v)$.