Drawing Representative Samples from Large Databases

Wen-Chi Hou
Southern Illinois University, USA

Hong Guo
Southern Illinois University, USA

Feng Yan
Williams Power, USA

Qiang Zhu
University of Michigan, USA

INTRODUCTION

Sampling has been used in areas like *selectivity* estimation (Hou & Ozsoyoglu, 1991; Haas & Swami, 1992, Jermaine, 2003; Lipton, Naughton & Schnerder, 1990; Wu, Agrawal, & Abbadi, 2001), OLAP (Acharya, Gibbons, & Poosala, 2000), clustering (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998; Palmer & Faloutsos, 2000), and spatial data mining (Xu, Ester, Kriegel, & Sander, 1998). Due to its importance, sampling has been incorporated into modern database systems.

The uniform random sampling has been used in various applications. However, it has also been criticized for its uniform treatment of objects that have non-uniform probability distributions. Consider the Gallup poll for a Federal election as an example. The *sample* is constructed by randomly selecting residences' telephone numbers. Unfortunately, the sample selected is not truly representative of the actual voters on the election. A major reason is that statistics have shown that most voters between ages 18 and 24 do not cast their ballots, while most senior citizens go to the poll-booths on Election Day. Since Gallup's sample does not take this into account, the survey could deviate substantially from the actual election results.

Finding *representative samples* is also important for many data mining tasks. For example, a carmaker may like to add desirable features in its new luxury car model. Since not all people are equally likely to buy the cars, only from a representative sample of potential luxury car buyers can most attractive features be revealed. Consider another example in deriving association rules from market basket data, recalling that the goal was to place items often purchased together in near locations. While serving ordinary customers, the store would like to pay some special tribute to customers who are handicapped, pregnant, elderly, and etcetera. A *uniform sampling* may not be able to include enough such under-populated people. However, by giving higher inclusion probabilities to (the transaction records of) these under-populated customers in sampling, the special care can be reflected in the association rules.

To find representative samples for populations with non-uniform probability distributions, some remedies, such as the density biased sampling (Palmer & Faloutsos, 2000) and the Acceptance/Rejection (AR) sampling (Olken, 1993), have been proposed. The density-biased sampling is specifically designed for applications where the probability of a group of objects is inversely proportional to its size. The AR sampling, based on the "acceptance/rejection" approach (Rubinstein, 1981), aims for all probability distributions and is probably the most general approach discussed in the database literature.

We are interested in finding a general, efficient, and accurate sampling method applicable to all probability distributions. In this research, we develop a Metropolis sampling method, based on the *Metropolis algorithm* (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), to draw representative samples. As it will be clear, the sample generated by this method is bona fide representative.

BACKGROUND

Being a representative sample, it must satisfy some criteria. First, the sample mean and variance must be good estimates of the population mean and variance, respectively, and converge to the latter when the sample size increases. In addition, a selected sample must have a

similar distribution to that of the underlying population. In the following, we briefly describe the population mean and variance and the Chi-square (Spiegel, 1991) test used to examine the similarity of distributions.

Mean and Variance Estimation

Let $\vec{x}$ be a d-dimensional vector representing a set of d attributes that characterizes an object in a population, and ρ the quantity of interest of the object, denoted as $\rho(\vec{x})$. Our task is to calculate the mean and variance of $\rho(\vec{x})$ of the population (relation). Let $w(\vec{x}) \geq 0$ be the probability or weigh function of $\vec{x}$. The population mean of $\rho(\vec{x})$, denoted by $\overline{\rho}$, is given by

$$\overline{\rho} = \sum_{all\ \vec{x}} \rho(\vec{x}) w(\vec{x}) \tag{1}$$

The probability distribution is required to satisfy

$$\sum_{all\ \vec{x}} w(\vec{x}) = 1. \tag{2}$$

For example, let $\vec{x}$ be a registered voter. Assuming there are only two candidates: a Republican candidate and a Democratic candidate, then we can let $\rho(\vec{x})$ = 1 if the voter $\vec{x}$ will cast a vote for the Republican candidate; and $\rho(\vec{x})$=-1, otherwise. $w(\vec{x})$ is the weight of the registered voter $\vec{x}$. If $\overline{\rho}$ is positive, the Republican candidate is predicted to win the election. Otherwise, the Democratic candidate wins the election.

Another useful quantity is the population variance, which is defined as

$$\Delta^2 = \sum_{all\ \vec{x}} (\rho(\vec{x}) - \overline{\rho})^2 w(\vec{x}). \tag{3}$$

The variance specifies the variability of the $\rho(\vec{x})$ values relative to $\overline{\rho}$.

Chi-Square Test

To compare the distributions of a sample and its population, we perform the Chi-square test (Press, Teukolsky, Vetterling, & Flannery, 1994) by calculating

$$\chi^2 = \sum_{i=1}^{k} (r_i - Nw_i)^2 / (Nw_i), \tag{4}$$

where r_i is the number of sample objects drawn from the i^{th} bin, $\sum_{i=1}^{k} r_i = N$ is the sample size, w_i is the probability of the i^{th} bin of the population, and Nw_i is the expected number of sample objects from that bin. A bin here refers to a designated group or range of values. The larger the χ^2 value, the greater is the discrepancy between the sample and the population distributions. Usually, a level of significance α is specified as the uncertainty of the test. If the value of χ^2 is less than $\chi^2_{1-\alpha}$, we are about 1-α confident that the sample and population have similar distributions: customarily, $\alpha = 0.05$. The value of $\chi^2_{1-\alpha}$ is also determined by the degree of freedom involved (i.e., k - 1).

MAIN THRUST

Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) has been known as the most successful and influential *Monte Carlo Method*. Unlike its use in numerical calculations, we shall use it to construct representative samples. In addition, we will also incorporate techniques for finding the best start sampling point in the algorithm, which can greatly improve the efficiency of the process.

Probability Distribution

The probability distribution $w(\vec{x})$ plays an important role in the Metropolis algorithm. Unfortunately, such information is usually unknown or difficult to obtain due to incompleteness or size of a population. However, the relative probability distribution or non-normalized probability distribution, denoted by $W(\vec{x})$, can often be obtained from, for example, preliminary analysis, knowledge, statistics, and etcetera. Take the Gallup poll for example. While it may be difficult or impossible to assign a weight (i.e., $w(\vec{x})$) to each individual voter, it can be easily known, for example, from Federal Election Commission, that the relative probabilities for people to vote on the Election Day (i.e., $W(\vec{x})$) are 18.5%, 38.7%, 56.5%, and 61.5% for groups whose ages fall in 18-24, 25-44, 45-65, and 65+, respectively. Fortunately, the relative prob-