A Model of Execution Time for the Bootstrap with PVM

Hiroyuki MINAMI and Masahiro MIZUTA
Center for Information and Multimedia Studies, Hokkaido University
N11 W5, Kita-ku,
Sapporo 060-0811, JAPAN
{*min,mizuta*}*@cims.hokudai.ac.jp*

1. Introduction

It is one of the fundamental topics for statistician to analyze huge data as fast as possible. A super computer would give us a solution but all of us cannot have this facility whenever we want. Some computer software libraries to make some personal computers one virtual parallel computer are available. We have implemented the Bootstrap algorithm according to Master-Slave model, suitable for parallel execution and investigate it through simulations.

In this paper, a model of execution time under the virtual machine by PVM which is one of the software libraries is proposed for the Bootstrap.

2. Parallel Virtual Machine(PVM) and Our Motivation

PVM(Geist, *et al.*, 1994) is a computer software library to make a virtual parallel computer from computers connected via network. This library is available on UNIX(and like) and MS WindowsNT/2000. We can use PVM environment with a program written in C or Fortran.

Most of statistical methods can formulate one master part which divides the data to calculate properly, send them to each slave and gather all results, and some slave parts which get data from the master and return results. This framework is suitable to implement an algorithm into a PVM machine.

We have studied various applications to utilize this framework. For example, Kawane *et al.* (2000) reported k-means method started with a lot of initial values simultaneously. In this paper, we focus on the Bootstrap.

3. Simulation

On the Bootstrap, the master part makes pseudo samples and the slaves calculate a statistic on each of samples. Here is an algorithm based on our concept.

1. Send bootstrap samples to all slaves
2. If some of slaves send the master a result, the master accepts it and send it another sample (until all samples are sent).

3. The master waits until all slaves return results.

In the paper, to make characteristics on execution time of our system clear, we implement a timer which can be set by 0.1 second in place of real calculation in slaves.

Now, we can construct our model of execution time as follows:

$$t_T = \alpha * N + \beta * B/N * t + \gamma + \varepsilon$$

$$\begin{cases} N\text{: Number of node (personal computer), } B\text{: Number of bootstrap samples} \\ t\text{: Parameter for a timer in slaves (second), } t_T\text{: Total execution time,} \\ \alpha, \beta, \gamma\text{: Constant, } \varepsilon\text{: Residual} \end{cases}$$

We set various conditions on N, B and t, and carried out the simulations 10 times for each condition.

Each personal computer has Intel Pentium3 450MHz CPU, 192MByte Memory and run under Redhat Linux 6.0. PVM Version of this simulation is 3.4.2.

On 300 data from the simulations, we can get the following model.

$$t_T = 0.0023 * N + 0.9971 * B * t/N - 0.0037$$

$$(R^2 = 0.99, \text{ Residual Standard Error} = 0.0125)$$

4. Discussion and Future Works

In this study, we can run our simulations under ideal environment. In reality, various factors would make effect the performance of PVM but we think our model is useful to consider N and execution time when one wants to do data analysis fast.

We report on the Bootstrap in the paper, however, our framework can be applied for many statistical methods. We are still studying this framework and have found some curious features, mainly related to the amount of data transfer between a master and a slave. We will report it soon.

REFERENCES

A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek & V. Sunderam(1994).

PVM: Parallel Virtual Machine, A Users' guide and Tutorial for Networked Parallel Computing, MIT Press.

M. Kawane, K. Komiya & M. Mizuta(2000). Cluster Analysis with Parallel Virtual Machine.

Proc. of 10th Japan and Korea Joint Conference of Statistics, 101–106.

H. Minami & M. Mizuta(2000). Empirical Study of Parallel Data Analysis.

Proc. of 10th Japan and Korea Joint Conference of Statistics, 131–136.

RÉSUMÉ

Dans ce papier, nous offrons un modèle pendant le temps d'exècution sur Bootstrap avec PVM. Nous examinons notre modèle par des exemples numériques.