An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?

Ryota Suzuki ryota.suzuki@is.titech.ac.jp

Hidetoshi Shimodaira shimo@is.titech.ac.jp

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama 2-12-1, Meguro-ku, Tokyo 152-8552, Japan

Keywords: microarray data, hierarchical clustering, multiscale bootstrap resampling, p-value

1 Introduction

The development of DNA microarray technology has enabled gene expression analysis based on simultaneous observation of thousands of genes. One of objectives of microarray analysis is to discover interesting structures from large amounts of expression data, and hierarchical clustering is often used to accomplish this. However, the estimation is often susceptible by statistical sampling error, and thus the result may be obtained only by chance without reflecting the true hypothesis. Therefore, it is necessary to evaluate the reliability of the hypothesis obtained from the analysis.

We applied multiscale bootstrap resampling (Shimodaira 2002 [4], 2004 [6]) to evaluate the accuracy of hypotheses as p-values. This method is based on resampling of data, and applicable to a large class of problems including hierarchical clustering. As an example, we show an application to hierarchical clustering of lung tumor microarray data.

2 Method

2.1 Hierarchical clustering

Cluster analysis is often used to bring similar individuals into groups. In hierarchical clustering, individuals are successively integrated based on the dissimilarity matrix computed by data, to obtain a dendrogram which contains inclusive clusters. In the context of microarray analysis, it is used to classify unknown genes or cases of disease.

2.2 Multiscale bootstrap resampling

In general, the result of hierarchical clustering for p individuals contains $p-1$ clusters. However, it is not clear how strong a cluster is supported by the data. We hope to know the accuracy of these clusters, where accuracy means the certainty of the existence of a cluster.

We measure the accuracy of these clusters as p-values, which ranges from 0 to 1. If the p-value of a cluster is less than α, the cluster is rejected at the α level of significance. Multiscale bootstrap resampling is a method which calculates p-values of hypotheses by resampling of data.

Note that the p-value calculated by multiscale bootstrap resampling is an approximation. But it is less biased than bootstrap probability, which is also an approximation of p-value computed by bootstrap resampling.

3 Example

We applied multiscale bootstrap resampling to the hierarchical clustering of microarray data of lung tumors in Garber et al (2001) [3]. The data is expression pattern of $n = 916$ genes of $p = 73$ tumors, and we conducted cluster analysis of tumors. We adopted $d(x_i, x_j) = 1 - \mathrm{cor}(x_i, x_j)$ as the dissimilarity matrix, and average method for clustering. In the multiscale bootstrap resampling, we took $n' = 467, 542, 636, 757, 916, 1131, 1431, 1869, 2544, 3664$ for the sample sizes of the bootstrap replicates, while the sample size of the original data is $n = 916$. This changes the scale of variation by the factor $\sqrt{n/n'}$. We generated B=10,000 replications for each scale value. Figure 1 shows the dendrogram computed by the original data, with p-values at some of branches.

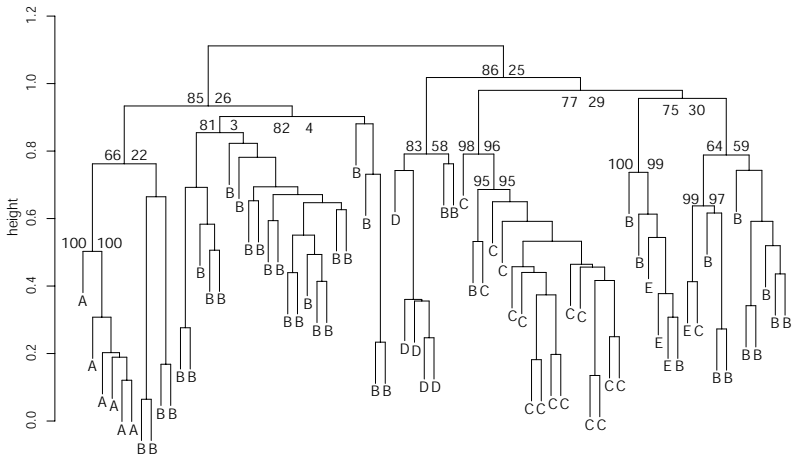

Figure 1: Hierarchical cluster analysis of 73 lung tumors. Values at branches are p-values (left), bootstrap probabilities (right) in percentage. Labels at leaves (A, B, C, D) are classification by specialists.

We developed a program for this analysis as an add-on package for a statistical package R [7]. It will be available at our website [8].

This work is supported in part by Grant-in-Aid for Young Scientists (A) KAKENHI-14702061 from MEXT of Japan.

References

[1] Efron, B., Bootstrap confidence intervals for a class of parametric problems, *Biometrika*, 72:45-58, 1985.

[2] Efron, B. and Tibshirani, R., The problem of regions, *Ann. Statist.*, 26:1687-1718, 1998.

[3] Garber, M. *et al.*, Diversity of gene expression in adenocarcinoma of the lung, *Proc. Natl. Acad. Sci. USA*, 98, 24:13784-13789, 2001.

[4] Shimodaira, H., An approximately unbiased test of phylogenetic tree selection, *Systematic Biology*, 51:492-508, 2002.

[5] Shimodaira, H., Assesing the uncertainty of the cluster analysis using the bootstrap resampling, Proceedings of the Institute of Statistical Mathematics, 80:33-44, 2002.

[6] Shimodaira, H. (in press), Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling, Annals of Statistics, 2004.

[7] `http://www.r-project.org/`

[8] `http://www.is.titech.ac.jp/~shimo/prog/`