BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 12 2006, pages 1540–1542
doi:10.1093/bioinformatics/btl117

Data and text mining

Pvclust: an R package for assessing the uncertainty in hierarchical clustering

Ryota Suzuki[1,2,*] and Hidetoshi Shimodaira[1]

[1]Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan and [2]Ef-prime, Inc., 2-17-5 Nihonbashi-Kayabacho, Chuo-ku, Tokyo 103-0025, Japan

Received on August 30, 2005; revised on March 6, 2006; accepted on March 25, 2006
Advance Access publication April 4, 2006
Associate Editor: Satoru Miyano

ABSTRACT

Summary: Pvclust is an add-on package for a statistical software R to assess the uncertainty in hierarchical cluster analysis. Pvclust can be used easily for general statistical problems, such as DNA microarray analysis, to perform the bootstrap analysis of clustering, which has been popular in phylogenetic analysis. Pvclust calculates probability values (*p*-values) for each cluster using bootstrap resampling techniques. Two types of *p*-values are available: approximately unbiased (AU) *p*-value and bootstrap probability (BP) value. Multiscale bootstrap resampling is used for the calculation of AU *p*-value, which has superiority in bias over BP value calculated by the ordinary bootstrap resampling. In addition the computation time can be enormously decreased with parallel computing option.

Availability: The program is freely distributed under GNU General Public License (GPL) and can directly be installed from CRAN (http://cran.r-project.org/), the official R package archive. The instruction and program source code are available at http://www.is.titech.ac.jp/~shimo/prog/pvclust

Contact: ryota.suzuki@is.titech.ac.jp

Cluster analysis is a statistical method which aims to classify several objects into some groups (clusters) according to similarities between them. While it has been widely used in many applications such as DNA microarray analysis, the uncertainty of results caused by sampling error of data has not generally been evaluated in practice. Pvclust is an implementation of bootstrap analysis on a statistical software R to assess the uncertainty in hierarchical cluster analysis.

The importance of uncertainty assessment has been well-recognized in phylogenetic analysis. It is a special form of hierarchical clustering for inferring the history of evolution as a dendrogram. Thousands of bootstrap samples are generated by randomly sampling elements of the data, and bootstrap replicates of the dendrogram are obtained by repeatedly applying the cluster analysis to them (Efron, 1979; Felsenstein, 1985). The bootstrap probability (BP) value of a cluster is the frequency that it appears in the bootstrap replicates. The multiscale bootstrap resampling was developed recently (Efron *et al*., 1996; Shimodaira, 2002, 2004) for calculating approximately unbiased (AU) probability values (*p*-values) as implemented in a software CONSEL (Shimodaira and Hasegawa, 2001).

*To whom correspondence should be addressed.

Although these bootstrap-based approaches are applicable to broad range of statistical problems, their usage have been limited since implementations of these methods are focused only on phylogenetic analysis. Pvclust is designed for general hierarchical clustering problems, so users can easily obtain bootstrap-based *p*-values for their own dataset and preferred clustering method. (Suzuki and Shimodaira, 2004).

R is 'a free software environment for statistical computing and graphics' (the R project website, http://www.r-project.org/), which runs on several platforms such as Windows, MacOS and UNIX/Linux. Several add-on packages are available via CRAN, the official R package archive.

Package pvclust is included in CRAN packages and can be easily installed on the fly. Once the package is installed, pvclust can be run with the command:

```
library (pvclust)
```

In R system data are stored in a 'data object'. For example, to read a data file `data.txt` into a data object `data`, type the following command:

```
data < - read.table(''data.txt'')
```

The bootstrap analysis is performed by applying the function `pvclust` to the object `data` with the number of bootstrap replications being $B = 10\,000$:

```
result < - pvclust(data, nboot=10000)
```

The result is stored in the object `result`, which can be shown graphically with the command:

```
plot(result)
```

We have applied the above commands to the DNA microarray data of Garber *et al*. (2001), and the result is shown in Figure 1. By default, `pvclust` performs bootstrapping at $K = 10$ different data sizes, and the hierarchical clustering is repeated by $K \times B$ times. It took 430 min on a single processor (Athlon MP 2000+). The parallel version `parPvclust` is available with the help of the snow package (Rossini *et al*., 2003), an add-on package of R for parallel computation. It took only 24 min using 20 processors; the speed-up value is 430/24 = 18, indicating very efficient parallel computing.

It is recommended to use `nboot = 1000` (default) for testing at first, followed by 10 000 for smaller errors. To determine an appropriate size of B (`nboot`), standard errors of *p*-values are helpful.

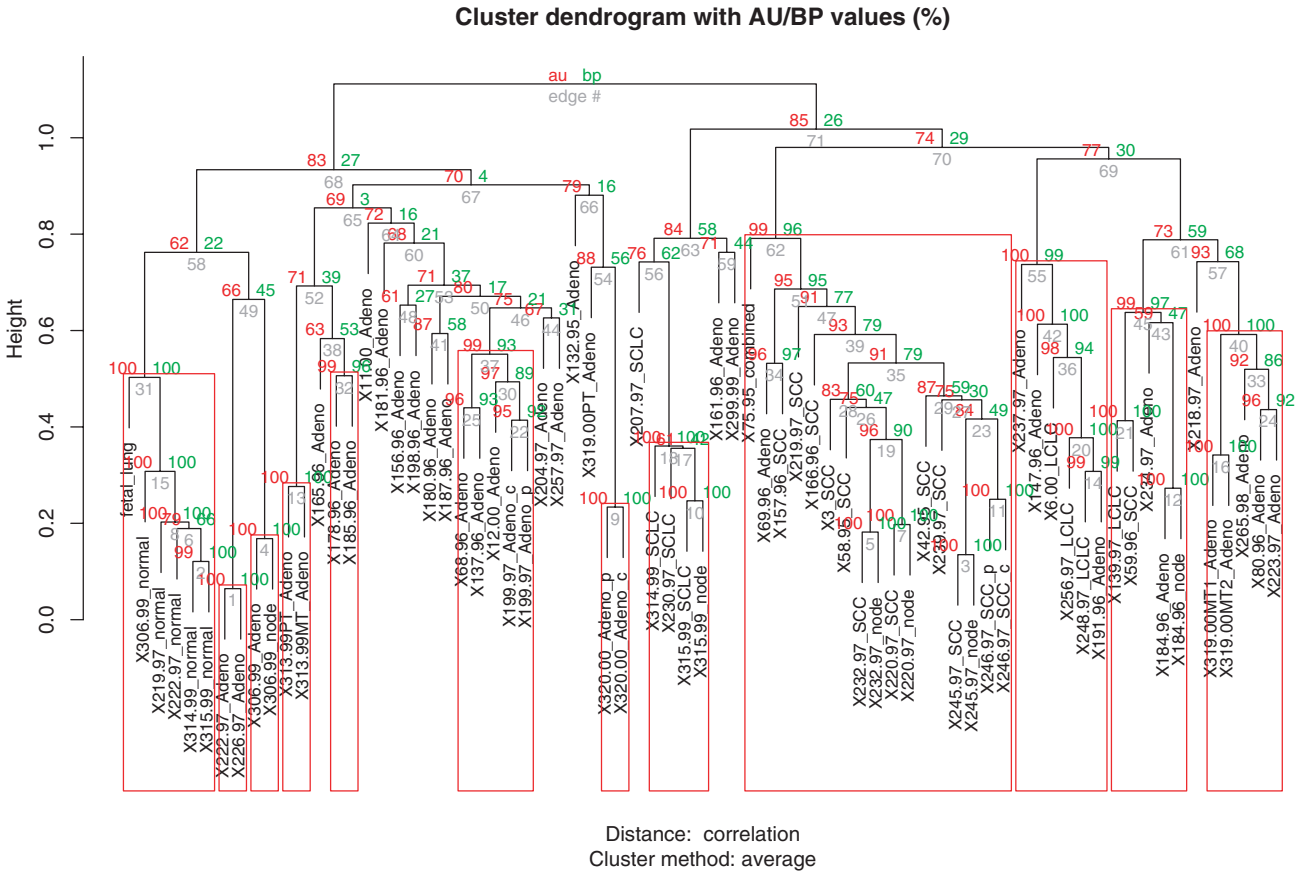

Fig. 1. Hierarchical clustering of 73 lung tumors. The data are expression pattern of 916 genes of Garber *et al*. (2001). Values at branches are AU *p*-values (left), BP values (right), and cluster labels (bottom). Clusters with AU ≥ 95 are indicated by the rectangles. The fourth rectangle from the right is a cluster labeled 62 with AU $= 0.99$ and BP $= 0.96$.

Function `seplot` provides a graphical interface for examining standard errors, while `print` gives more detailed information about *p*-values in text-based format. See online instruction on our website for the usage of these facilities.

In the multiscale bootstrap resampling, we intentionally alter the data size of bootstrap samples to several values. Let N be the original data size, and N' be that for bootstrap samples. In the example of Figure 1, $N = 916$, and $N' = 458$, 549, 641, 732, 824, 916, 1007, 1099, 1190 and 1282. For each cluster, an observed BP value is obtained for each value of N', and we look at change in $z = -\Phi^{-1}$ (BP) values, where $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$, the standard normal distribution function. For the cluster labeled 62 in Figure 1, the observed BP values are 0.8554, 0.8896, 0.9132, 0.9335, 0.9498, 0.9636, 0.9656, 0.9756, 0.9795 and 0.9859 (Fig. 2). Then, a theoretical curve $z(N') = v\sqrt{N'/N} + c\sqrt{N/N'}$ is fitted to the observed values, and the coefficients v, c are estimated for each cluster. The AU *p*-value is computed by AU $= \Phi(-v + c)$. For the cluster labeled 62, $v = -2.01$, $c = 0.26$, and thus AU $= \Phi(2.01 + 0.26) = \Phi(2.27) = 0.988$, where BP $= 0.964$ for $N' = N$. An asymptotic theory proves that the AU *p*-value is less biased than the BP value.

Currently only the simplest form of the bootstrapping, i.e. the non-parametric bootstrap resampling, is implemented in pvclust. More elaborate models designed for specific applications, such

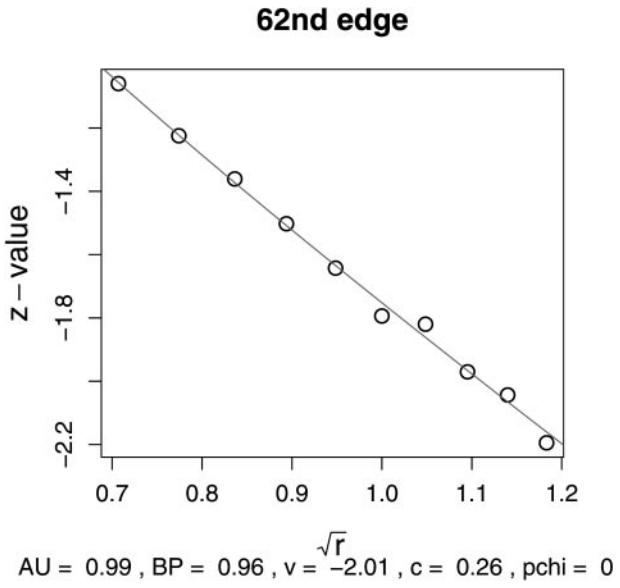

Fig. 2. Diagnostic plot of the multiscale bootstrap for the cluster labeled 62. The observed *z*-values are plotted for $\sqrt{N'/N}$, and the theoretical curve is obtained by the weighted least squares fitting. This plot is obtained by command: `msplot(result, edges=62)`. When the curve fitting is poor, a breakdown of the asymptotic theory may be suspected.

as that of Kerr and Churchill (2001) for DNA microarray analysis, should be incorporated into the program in a future work.

ACKNOWLEDGEMENTS

This work is supported in part by Grant KAKENHI (14702061, 17700276) from MEXT of Japan.

Conflict of Interest: none declared.

REFERENCES

Efron,B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

Efron,B. *et al.* (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci., USA*, **93**, 13429–13434.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Garber,M. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. [Erratum (2002) *Proc. Natl Acad. Sci. USA,* 99, 1098.] *Proc. Natl Acad. Sci. USA*, **98**, 13784–13789.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Rossini,A. *et al.* (2003) Simple parallel statistical computing in R. *UW Biostatistics Working Paper Series. Paper 193*, University of Washington, WA.

Shimodaira,H. and Hasegawa,M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**, 1246–1247.

Shimodaira,H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.*, **51**, 492–508.

Shimodaira,H. (2004) Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann. Stat.*, **32**, 2616–2641.

Suzuki,R. and Shimodaira,H. (2004) An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: how accurate are these clusters? In *proceedings by the Fifteenth International Conference on Genome Informatics (GIW 2004)*, p. P034.