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

How accurate are these clusters?

Cluster analysis is a method to examine the similarities between individuals. Hierarchical clustering generates a dendrogram which contains clusters which show such similarities based on the dissimilarity matrix computed by data. It offers detailed information on the relationships between individuals and hence has been often used in applied areas including microarray analysis. However, it is not clear how strong these clusters are supported by the data. The question is, "How accurate are these clusters?"

Given data and an algorithm to construct a dendrogram, hierarchical clustering generates a dendrogram which contains clusters, as shown in Figure 1. We have microarray data of sample size n = 10, where columns correspond to p = 4 individuals to be analyzed and rows correspond to n = 10 observations. By the resulting dendrogram, we hope to conclude, for example, that A and B are closer than B and C. If we can take samples as many as we want, we can examine this conclusion by applying the same analysis to these samples, as shown in Figure 2.

However, as is always the case, we have only one sample available. Hence we cannot deny the possibility of the situation shown in Figure 3. The figure shows the space of data divided by the resulting dendrograms, where our sample is indicated by the red star. The sample is drawn from the probability distribution centered at the blue star, which is just inside the area where B and C are closer than any other combination of two individuals. In this situation our conclusion is obtained because of random noise and does not correctly reflect the truth.

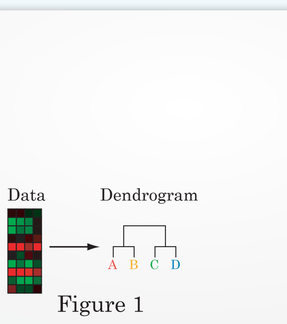

Figure 1

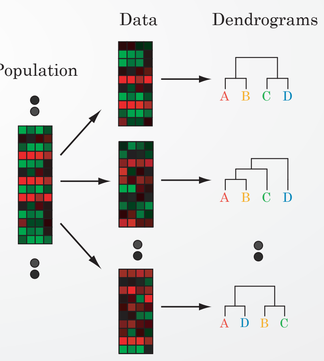

Figure 2

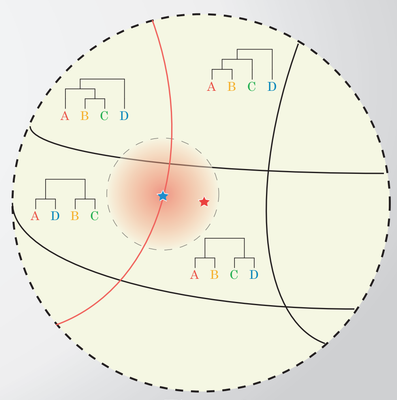

Figure 3

Bootstrap probability

How can we assess such uncertainty using only one sample? One way of achieving this is using bootstrap resampling. In bootstrap resampling we replicate data by resampling from the data itself. Figure 4 shows this situation. We randomly choose n = 10 observations from the original data with replication. The procedure is repeated many times, as 10,000 times in the figure. These sampled are called bootstrap samples.

We can examine the result of hierarchical clustering using bootstrap samples, through quantities called "bootstrap probabilities". Figure 5 shows the procedure in detail. First, we generate bootstrap samples. Second, we apply hierarchical clustering to each bootstrap sample. The resulting dendrograms are called bootstrap replications of dendrograms. Third, we count the number of dendrograms which contained a cluster which we hope to examine. In this example, we took the cluster (B, C), which is called a hypothesis. Finally, divide this number by the number of bootstrap samples to obtain the ratio of the dendrograms which fulfill the hypothesis.

The ratio is called bootstrap probability and it can be used to examine the certainty of the hypothesis. In this example, the bootstrap probability of the cluster (B, C) was 4%. It is quite small and it seems that we may be able to conclude that the cluster (B, C) does not exist in the true dendrogram. Indeed this conclusion is correct in an approximate sense. However the approximation is not enough good, so we will adopt a more sophisticated way.

Figure 6 shows what bootstrap probability means. In bootstrap resampling we generate a probability distribution centered at the original data, indicated as the red star. The bootstrap probability we computed is the probability that the bootstrap samples fall inside the area where B and C are most closer. It reflects the uncertainty of cluster analysis in some seanses, but this does not satisfy our objective since it is not exactly the same situation we hope to know, which is shown in Figure 3. The center of the distribution should not be the same point as our sample, but it should be inside the area of our hypothesis.

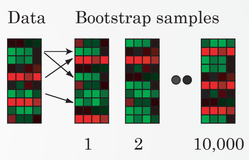

Figure 4

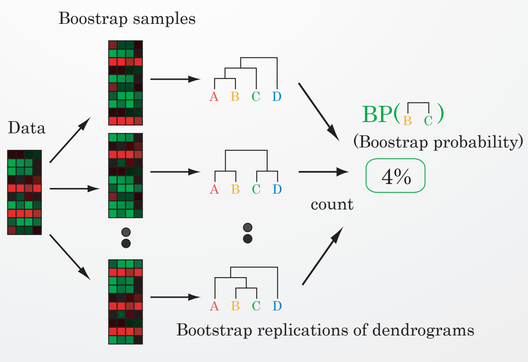

Figure 5

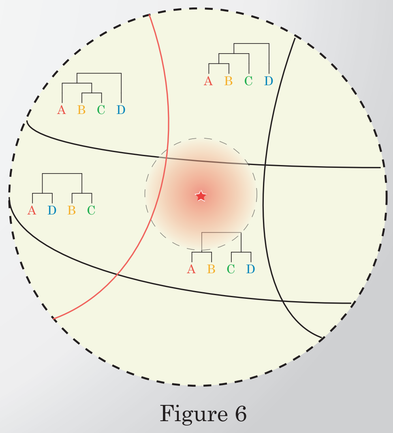

Figure 6

Multiscale bootstrap resampling

By comparing Figure 3 and 4, we can see that there is a gap between our objective and bootstrap probability. Multiscale bootstrap resampling [2] successfully fills this gap. In bootstrap resampling, the sample size of a bootstrap sample was n, the same as that of original data. On the other hand, we change the sample sizes to several values in multiscale bootstrap resampling. We take sample sizes n', which can be smaller than or larger than, or also can be equal to the original size n. It is shown in Figure 7. In the figure original sample size is n = 10, and bootstrap samples with n' = 5, 10, 15 are shown.

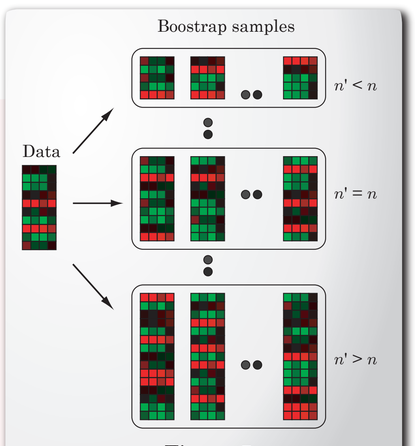

Figure 7

Using these bootstrap samples, multiscale bootstrap resampling computes a quantity called p-value for each hypothesis. If the p-value of a hypothesis is very small, say smaller than 5%, we can reject the hypothesis. In fact bootstrap probability is an approximation of this value, and multiscale bootstrap resampling corrects the bias of bootstrap probability.

The algorithm of multiscale bootstrap is shown in Figure 8. First, we generate bootstrap samples for each sample size. Second, we apply hierarchical clustering to each bootstrap sample to obtain the sets of bootstrap replications of dendrograms. Third, we compute bootstrap probability for each sample size. Finally, using values of bootstrap probabilities, we can estimate the p-value by fitting a theoretical equation to them. The estimated p-value is called AU (approximately unbiased) value.

Figure 9 shows what multiscale bootstrap resampling computes. AU value is equivalent to the probability that a new sample (if available) appears farther than our sample, under the given hypothesis. In the figure, the red star is our sample and the blue star is the center of the probability distribution, just inside the area of hypothesis. The red dashed curve indicate the points where the distance between data and the hypothesis is the same as our sample. AU value is the probability that a new sample is outside the area indicated by the red dashed curve. If this probability is less than, for example 5%, we can conclude that our sample is not obtained under the hypothesis. In this example AU value is 8%, so we cannot deny the possibility that the data is obtained under the hypothesis that B and C are most closer.

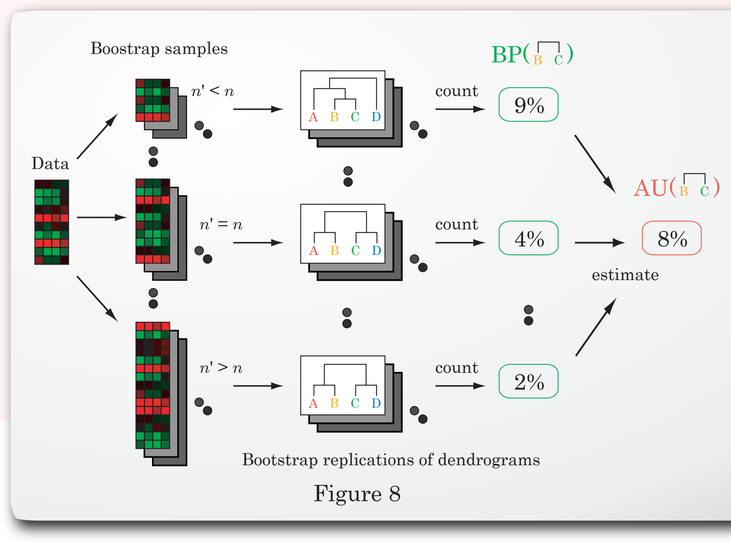

Figure 8

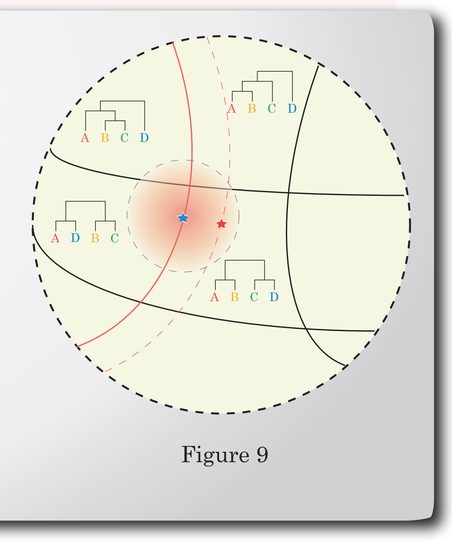

Figure 9

Example

We applied multiscale bootstrap resampling to the hierarchical clustering of microarray data of lung tumores, in Garber et al [1]. The data is expression pattern of n = 916 genes of p = 73 tumores, and we conducted cluster analysis of tumores. We took n' = 467, 542, 636, 757, 916, 1131, 1431, 1869, 2544 and 3664 for multiscale bootstrap resampling, and generated B = 10,000 replications for each sample size. Figure 10 shows the result of this analysis. The labels from A to E are classification by specialists and it can be seen that cluster analysis returns a similar result. We can also see that some clusters have quite high AU values as larger than 95%. In these cases, we can reject the hypothesis that these cluster do not exist. In other words, we can conclude that these clusters are strongly supported by the data. In this example, the cluster which contains all A's is such a case. Actually the label A means normal cell and cluster analysis seems to detect this strong feature in the data.

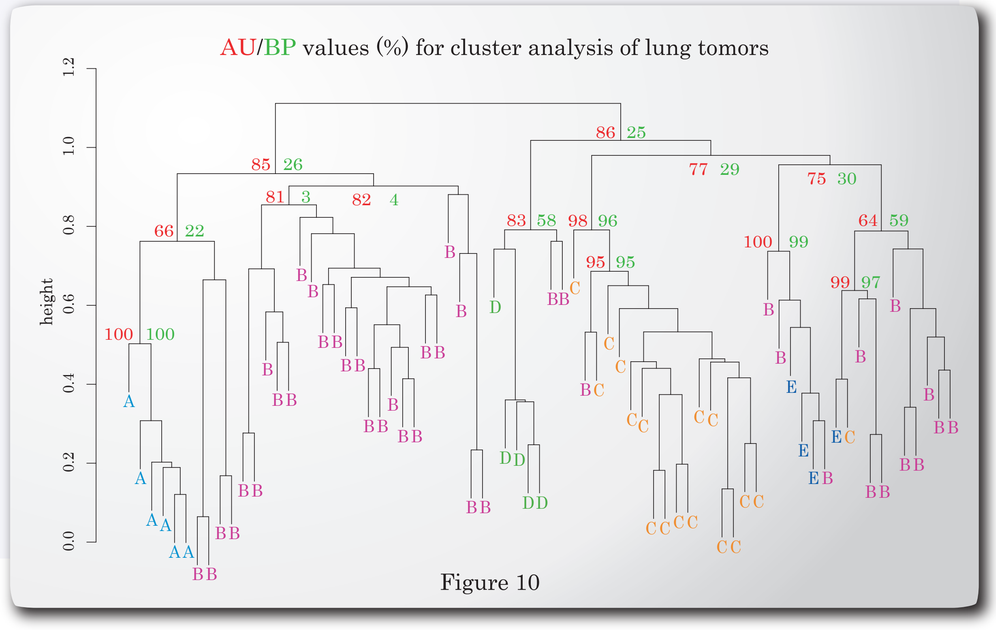

Figure 10

References

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

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