`sigclust2`: Statistical Significance of Clustering (..part 2)

Patrick K. Kimes

September 18, 2014

Contents

1 Summary

sigclust2 includes significance tests for the output of various popular clustering algorithms. The `sigclust2` approach is conceptually based on the significance testing procedure originally described in [1] and implemented in *sigclust*. While *sigclust* was only capable of testing for the significance of two clusters, we generalize the testing approach to a larger number of clusters, as well as the (agglomerative) hierarchical clustering procedure.

2 Introduction

Hierarchical clustering is a popular unsupervised learning tool commonly implemented in genomics to identify subgroup behavior in unlabeled data sets, e.g. a cohort of tumor samples. Unlike non-nested/flat clustering, hierarchical clustering does not require pre-specifying the number of subgroups of interest. Furthermore, hierarchical clustering provides an intuitive representation of the clustering results as a binary tree (*dendrogram*). In cluster analysis, a question of substantial interest and importance is that of whether subgroups identified in the data truly exist. In this package, we present Significance for Hierarchical Clustering (SHC, `sch`), a simulation based hypothesis testing procedure for identifiying significant clustering in a hierarchical cluster.

In this tutorial we describe our algorithm using one of the simulation settings described in the accompanying manuscript. We also provide steps for reproducing the simulation results presented in the manuscript.

2.1 Citation

For a more in depth discussion of the apporach including the multiple testing correction procedure implemented in *sigclust2*, contact Patrick. The corresponding manuscript is still in preparation.

> Kimes, P. K., Hayes, D. N., Liu Y., and Marron, J. S. (2014) Statistical significance for hierarchical clustering.
> *in preparation*.

3 `shc` Implementation

```
# library(sigclust2)
# generate some data
# apply the method to the data
# reader: "wow! that was so easy!"
```

4 Visualization

The results of *sigclust* may be summarized in a dendrogram.

```
# plot the results of the implementation from above
```

5 Full Simulation

In this section we show how the code from the previous sections were combined to obtain the simulation results presented in the accompanying manuscript.

References

[1] Liu, Y., Hayes, D. N., Nobel, A., Marron, J. S. (2008) Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data. Journal of the American Statistical Association 103(408), 1281—1293.

[2] The TCGA Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature 489(7417), 519—525.