Robust Model-Based Clustering of Genes in Microarray Data: Are there Gene Clusters?

Johanna Hardin
Fred Hutchinson Cancer Research Center
1100 Fairview Ave N MP-557 P.O. Box 19024 Seattle, WA 98109-1024
USA
(206) 667-4846
johannah@swog.fhcrc.org
CAMDA00 Dataset 2: Leukemia

Johanna Hardin, Southwest Oncology Group, Fred Hutchinson Cancer Research Center David M. Rocke, University of California, Davis David L. Woodruff, University of California, Davis

Microarray technology has produced the capability to collect data on large amounts of genetic information. We provide a method of refinement to the clustering of genes based on such technology with the idea that these clusters might lead to the discovery of genetic pathways that cause various diseases. Our method uses Rousseeuw's Minimum Covariance Determinant (MCD) (Lopuhaa and Rousseeuw, 1991; Rousseeuw and Leroy, 1987) as a robust measure of location and shape. We use the MCD of each cluster as a robust, data-dependent, measure of each of the different clusters. We then apply F-distribution quantiles to the Mahalanobis squared distances of the data, based on the MCD parameters, to find points which may in fact belong to more than one cluster or may be outliers which do not seem to belong to any cluster (Hardin and Rocke, 2001a; Hardin and Rocke, 2001b). The results are applied to the leukemia data of Golub et al. (Golub et at., 1999) after subsetting for genes that are primarily expressed as present across the samples. The clusters are then analyzed as to their validity and usefulness.

Keywords

Robust Clustering, Outliers, Minimum Covariance Determinant, EMMIX

Tools

The clustering software we used as an initialization is called EMMIX and is available at:
http://www.maths.uq.oz.au/~gjm/emmix/emmix.html

website

No website available