Consistent biclustering by sparse singular value decomposition incorporating stability selection

Martin Sill and Axel Benner
Division of Biostatistics, German Cancer Research Center
Im Neuenheimer Feld 280, Heidelberg, Germany
m.sill@dkfz.de
benner@dkfz.de

1.07.2010

Abstract

High-dimensional gene expression data arises in all fields of life science and is usually stored as two-way two-mode data matrix. Often interest lies in finding a set of genes that show a correlated gene expression within a subset of the samples. In order to find solutions to this two-way clustering problem a large number of so called biclustering approaches have been proposed. Most of these algorithms aim to find biclusters that correspond to submatrices of the gene expression matrix.

In an idealized case, e.g. assuming low noise and assuming that the gene expression matrix has a block-diagonal structure, each block displays a unique bicluster. Decomposing this idealized data matrix by singular value decomposition (SVD) results in a matrix factorization where each singular vector pair is associated with a bicluster. The nonzero elements of the left and right singular vectors are the gene and sample subsets defining the bicluster. Therefore many of the existing biclustering methods are strongly related to SVD, e.g. the Plaid Model and the Iterative Signature Algorithm. But applying SVD in case of non-idealized real data sets will result in singular vectors with a large number of non zero elements. Recently, a sparse SVD method has been proposed and successfully applied to find reasonable biclusters in gene expression data. This regularized form of the SVD alternately fits penalized regression models to the singular vector pair to obtain a sparse matrix decomposition. The resulting sparse singular vectors strongly depend on the choice of the right penalization parameter.

We propose to choose the right amount of penalization by incorporating a stability selection. The stability selection is a subsampling procedure that can be applied to penalized regression models to find stable variables and additionally offers the possibility of an error control. The practical application of this new biclustering approach is demonstrated by an analysis of gene expression data sampled from different ependymoma (brain and spinal tumors) subtypes.

References

BERGMANN, S., IHMELS, J. and BARKAI, N. (2003): Iterative signature algorithm for the analysis of large-scale gene expression data *P*hysical Review (E67) ,031902

BUSYGIN, S., PROKOPYEV, O. and PARDALOS P.M. (2008): Biclustering in data mining *Computers & Operations Research (35) ,2964-2987*

LAZZERONI, L. and OWEN, A. (2002): Plaid Models for gene expression data *S*tatistica Sinica, 61-86

LEE, M., SHEN, H., HUANG, J.Z. and MARRON, J.S. (2010): Biclustering via Sparse Singular Value Decomposition *Biometrics, DOI 10.1111/j.1541-0420.2010.01392.x*

MEINSHAUSEN, N. and BÜHLMANN, P. (2009): Stability Selection *Preprints of Journal of the Royal Statistical Society, To be published in Series B*

VAN MECHELEN, I., BOCK,H.H. and DE BOECK P. (2004): Two-mode clustering methods: a structured overview *S*tatistical Methods in Medical Research (13) 363–394