JENA CENTRE FOR BIOINFORMATICS

SPONSORED BY THE
Federal Ministry of Education and Research

Advantages of a global optimization approach for the cluster analysis of gene expression data

Ulrich Möller, Frank Thies
UMoeller@pmail.hki-jena.de, FThies@pmail.hki-jena.de

Hans Knöll Institute for Natural Products Research (HKI)
Beutenbergstrasse 11a, D-07745 Jena; Germany

Problem

Cluster analysis based on function optimization schemes has proven useful for the characterization of unknown structure in the field of bioinformatics. However, the result of a given algorithm may depend on chance due to random initialization or a stochastic optimization rule, and heuristic settings. In several studies, rerunning an algorithm on the same data set yielded quite different partitions from which analysts would derive ambiguous conclusions [1,2]. This problem potentially prevents an appropriate interpretation of the biological aspects associated with the clustering result of a data set.

Goals

- quantifying the variability of clustering results of the same algorithm
- demonstrating the potential divergence of the clustering partitions
- making aware of a conceptual framework for the evaluation of a given algorithm's capability that is independent of heuristics and randomness
- comparing the effectiveness and efficiency of several clustering tools

Methods

Partitioning cluster analysis using several types of calculus

- K-means (KM) and fuzzy K-means (FKM) algorithm (see e.g. [3])
- random search among cluster centroids (RSC) [4]
- stochastic relaxation with decoder perturbation (SRD) [5]

Statistical evaluation of an algorithm [4]

- rerunning each algorithm multiple times involving a different initialization and/or a different course of stochastic optimization
- quantification of effectiveness and efficiency parameters
- graphical representation of partitions at different optimization levels

Data

Yeast cell cycle gene expression data [6] (http://genomics.stanford.edu)
Data preprocessing according to [7]:

A **variation filter** was used to eliminate those genes that did not show significant changes during the time course: i) an absolute value of expression at all 17 time points of equal to or greater than 100 (in units in the downloaded file); ii) at least a 2.5-fold change in expression level during the time course. 1306 out of more than 6000 gene expression patterns passed the variation filter. These data were **normalized** such that the expression level varied between 0 and 1.

Discussion

The recognition of particular, biologically relevant clusters of genes may require that a high level of optimization has been achieved for the objective function of clustering. This optimization level can be reached more efficiently when using clustering algorithms based on a global rather than a local optimization strategy.
Statistical evaluation of multiple attempts of clustering is a useful method in order to provide evidence that the optimal, or a near-optimal, partition has been found.

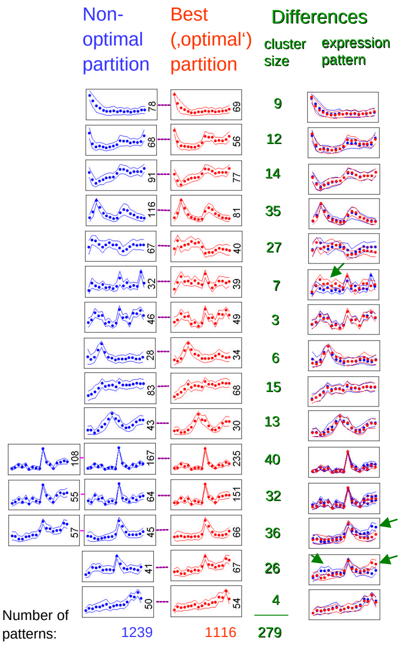
Literature

[1] Möller et al.: Verbesserte Strukturierung von vorsegmentierten EEG-Abschnitten durch Clusterbildung mit globaler Optimierung. Eine methodische Studie. Z. EEG-EMG 27 (1996) 105-110
[2] Möller et al.: Pitfalls in the clustering of neuroimage data and improvements by global optimization strategies. NeuroImage 14 (2001) 206-218
[3] Theodoridis and Koutroumbas: Pattern Recognition, Academic Press, San Diego, 1998
[4] Möller et al.: An efficient vector quantizer providing globally optimal solutions. IEEE Transactions on Signal Processing 46 (1998) 2515-2529
[5] Zeger et al.: Globally optimal vector quantizer design by stochastic relaxation. IEEE Transactions on Signal Processing 40 (1992) 310-322
[6] Cho et al.: A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2 (1998) 65-73
[7] Lukashin and Fuchs: Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics 17 (2001) 405-414

Results (selection)

Clustering model	Hard			Fuzzy
Algorithm	**RSC**	**SRD**	**KM**	**FKM**
Minimum error [1]	**0.00**	0.01	0.63	0.00
Average error [1]	0.62	0.59	2.47	0.04
Maximum error [1]	1.73	1.94	8.45	0.37
Iterations (average)	178.8	178	26.2	28.1
Iterations (best trial)	170	178	44	41
Iterations (longest trial)	239	178	67	73
Computation time (trial) [2]	≈ 18 s	≈ 19 s	≈ 2 s	≈ 11 s

[1] Each algorithm was rerun 1000 times with a random initialization and/or a random search path in each run (trial). The smallest value of the objective function (of hard or fuzzy clustering) was set to 1 (reference value). The minimum, average and maximum errors denote the percentage by which the objective function values of the 1000 trials were larger than the reference. The objective function was the sum of squared Euclidean distances between the cluster members and their cluster center.
[2] Pentium 4, 2.2 GHz, 2 GB RAM, C program, Linux environment



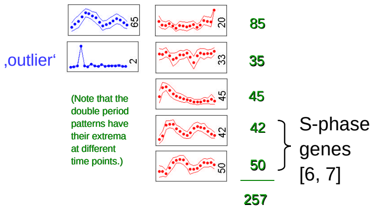
The number of clusters was set K = 20 according to the results in [7].

The best K-means result, with respect to the objective function, was only as good as the RSC result on average, where K-means required a total of 26,200 iterations (see above table).

Each algorithm provided a number of different partitions of which some clusters were non-comparable. The two partitions on the left-hand side are one example.

536 of 1306 patterns (41%) were assigned to different clusters/ classes.

┈ These clusters roughly correspond to each other with respect to their expression pattern.



(Note that the double period patterns have their extrema at different time points.)

Some clusters that were obtained in the the empirical optimum solution, and not in the sub-optimal solution, represent patterns of biologically characterized genes.

The red clusters do not have an equivalent expression in the non-optimal (blue) partition.

The JCB is a member of the NBCC.