

Assessing the stability of unsupervised learning results in small-sample-size problems

Ulrich Möller

Email: Ulrich.Moeller@hki-jena.de



**Leibniz Institute for Natural Product Research
and Infection Biology – Hans Knöll Institute**



Jena Center for Bioinformatics

GMDS-Tagung 2006

Bioinformatik 1

**Stabilitätsanalyse
von Ergebnissen des unüberwachten Lernens
aus kleinen Stichproben**

Ulrich Möller

Email: Ulrich.Moeller@hki-jena.de



**Leibniz-Institut für Naturstoff-Forschung und
Infektionsbiologie – Hans-Knöll-Institut**



Jena Center for Bioinformatics

Gliederung



Einleitung

- Clustering von Microarray-Daten
- Validierung mittels Resampling

Leistungsvergleich von Resampling-Techniken

- Strategie, Methoden, Daten
- Hauptergebnis

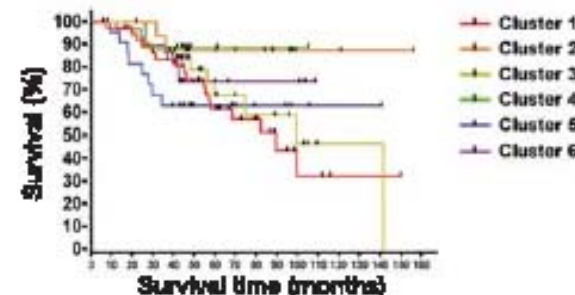
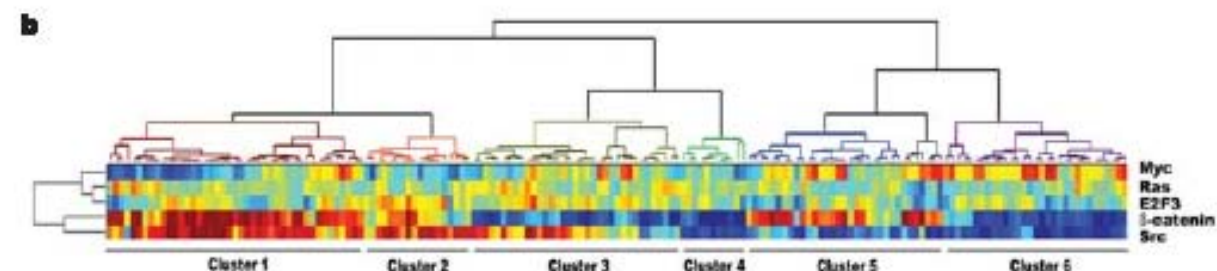
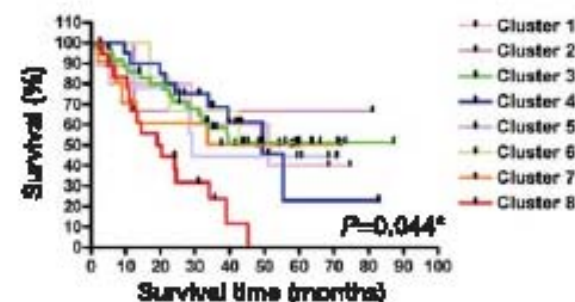
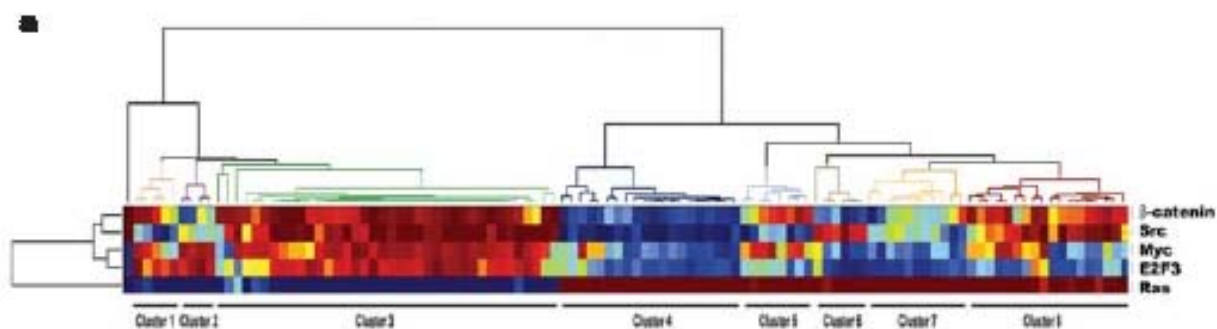
Nächste-Nachbarn-Resampling

- Idee
- Ergebnisse

Schlussfolgerungen

Oncogenic pathway signatures in human cancers as a guide to targeted therapies

Andrea H. Bild^{1,2}, Guang Yao^{1,2}, Jeffrey T. Chang^{1,2}, Quanli Wang¹, Anil Potti^{1,4}, Dawn Chasse^{1,2}, Mary-Beth Joshi³, David Harpole³, Johnathan M. Lancaster⁷, Andrew Berchuck⁵, John A. Olson Jr^{1,3}, Jeffrey R. Marks³, Holly K. Dressman^{1,2}, Mike West⁶ & Joseph R. Nevins^{1,2}



Microarray-Datencluster

Clustervalidierung

BIOINFORMATICS

REVIEW

Vol. 21 no. 15 2005, pages 3201–3212
doi:10.1093/bioinformatics/bti517

Data and text mining

Computational cluster validation in post-genomic data analysis

Julia Handl*, Joshua Knowles and Douglas B. Kell

School of Chemistry, University of Manchester, Faraday Building, Sackville Street, PO Box 88,
Manchester M60 1QD, UK

Received on March 24, 2005; revised and accepted on May 24, 2005

Advance Access publication May 24, 2005

Liefert ein Clusteralgorithmus

für Stichproben aus **derselben** (Misch-)Population

eine **stabile** Partitionierung ?

Resampling – auf welche Weise ?



www.wuerfel-offensive.de

Bootstrapping
(parametrisch)

Addition von
Rauschen
(parametrisch)

Subsampling

Bootstrapping
(nicht-parametrisch)

Addition von
Rauschen
(nicht-parametrisch)


weitere

Gliederung

Einleitung

- Clustering von Microarray-Daten
- Validierung mittels Resampling

Leistungsvergleich von Resampling-Techniken

- 
- Strategie, Methoden, Daten
 - Hauptergebnis

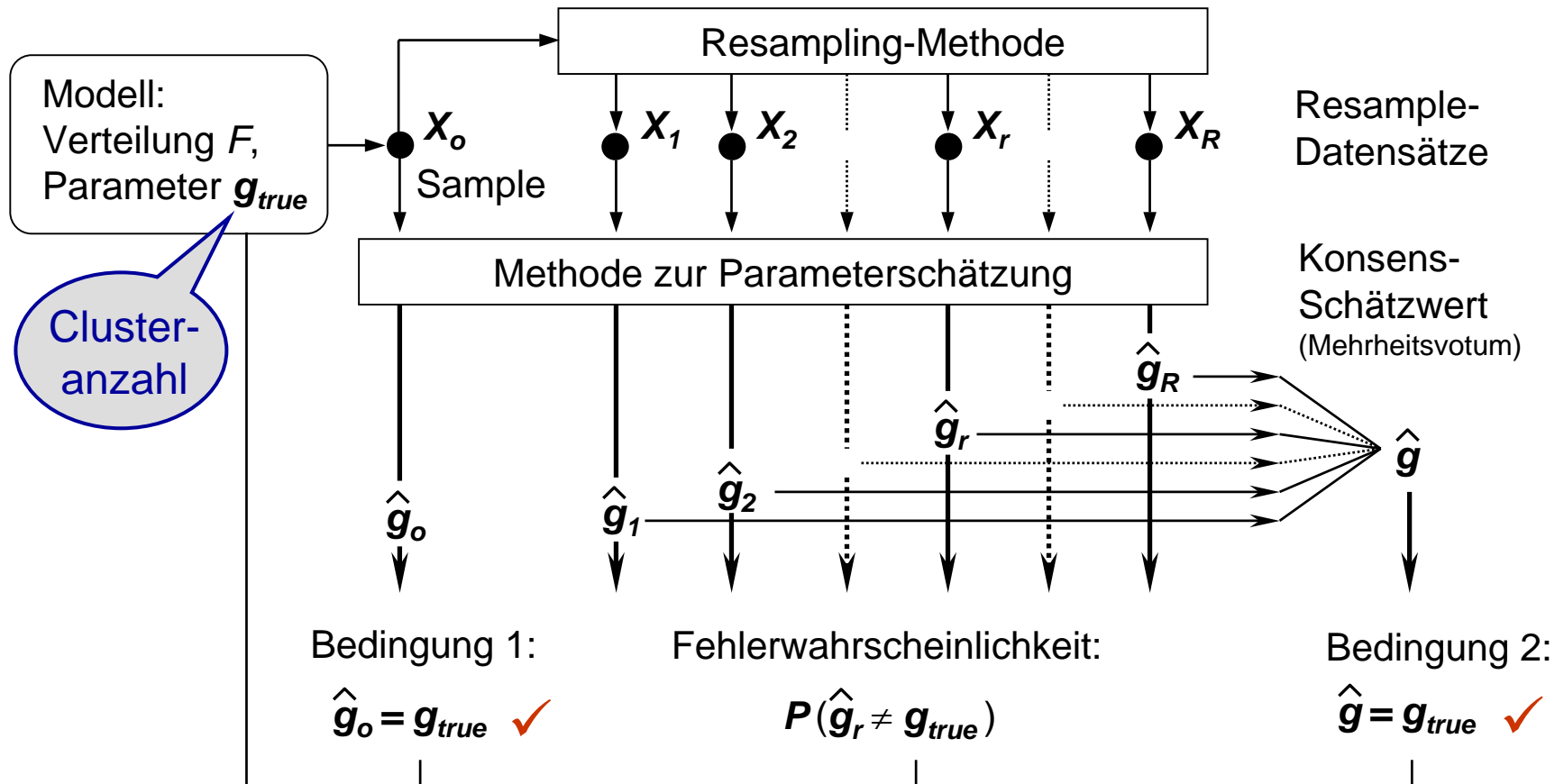
Nächste-Nachbarn-Resampling

- Idee
- Ergebnisse

Schlussfolgerungen

Leistungsvergleich von Resampling-Techniken

Strategie



Leistungsvergleich von Resampling-Techniken

Methoden zur Schätzung der Clusteranzahl

Estimator of the number of clusters (validity index)	Method	Hierarchical clustering ①⑥	K-medoid clustering MEDOID ②	Fuzzy C-means clustering FCM1 ⑥	Fuzzy C-means re-clustering FCM50 ③⑥	Total number
		no.	1 2 3 4 5 6 7 8 9	10	11	
Generalized Davis-Bouldin Indices DBI [4, 7] ④	1 10 ⋮ ⋮ 9 18					216
Generalized Dunn Indices DI [4] ④	19 28 ⋮ ⋮ 27 36					216
Mean silhouette width MSI [24] ⑥	37					12
Global stability SI [18]	38					11
PI (Pakhira) [35]	39					1
RI (Ray) [38]	40					1
FI (Fadili) [17]	41					1
Total number		342	38	38	40	458

Leistungsvergleich von Resampling-Techniken

Daten I: Stochastische Modelle

Name	p	n_i	C_{true}	\mathbf{X}_{ij} ($j = 1, \dots, n_i$)
model 2	2	25, 25, 50	3	$\mathbf{X}_{ij} \sim N(\mu_i, \mathbf{I}_2)$, $\mu_1 = (0, 0)^T$, $\mu_2 = (0, 5)^T$, $\mu_3 = (5, -3)^T$
model 3	10	25 or 50 ①	4	$\mathbf{X}_{ij} \sim N(\mu_i, \mathbf{I}_{10})$, $\mu_i = (\mathbf{w}_i^T, \mathbf{0}_7^T)^T$, $\mathbf{W}_i \sim N(\mathbf{0}_3, 25\mathbf{I}_3)$ ②
model 4	10	25 or 50 ①	4	$\mathbf{X}_{ij} \sim N(\mu_i, \mathbf{I}_{10})$, $\mu_i = \mathbf{w}_i$, $\mathbf{W}_i \sim N(\mathbf{0}_{10}, 3.6\mathbf{I}_{10})$ ②
model 6	10	100	2	$\mathbf{X}_{ij} = (\mathbf{X}_{1ij}^T, \mathbf{X}_{2ij}^T)^T$, $\mathbf{X}_{1ij} \sim N(\mu_{1j}, \mathbf{I}_3)$, $\mathbf{X}_{2ij} \sim N(\mathbf{0}_7, \mathbf{D}_7)$ ③ $\mu_{1j} = -0.5 + 0.1(j-1)/99$, $\mu_{2j} = \mu_{1j} + 10$
model 7	10	50	2	$\mathbf{X}_{ij} \sim N(\mu_i, \mathbf{I}_{10})$, $\mu_1 = \mathbf{0}_{10}$, $\mu_2 = (2.5, \mathbf{0}_9^T)^T$

p number of variables (dimensions) \mathbf{X}_{ij} simulated observation j of cluster i \mathbf{I}_p $p \times p$ identity matrix

n_i cluster sizes \mathbf{w}_i realization of the random variable \mathbf{W}_i T transpose

① with probability 0.5 each

② only simulations where any two members of different clusters have a minimum Euclidean distance ≥ 1

③ \mathbf{D}_7 : diagonal matrix with the elements $d_{i,i} = (i+3)^2$, $i = 1, \dots, 7$

Dudoit S, Fridlyand J, Genome Biology, 2002

U. Möller, D. Radke: Performance of data resampling methods for robust class discovery based on clustering. Intelligent Data Analysis 10(2), 2006, 139-162

Leistungsvergleich von Resampling-Techniken

Daten II: Genexpressionsdaten (Tumor)

Name	p	n	C_{biol}	Number of samples with biological characterization based on the phenotype
leukemia	200	38	3	11 acute myeloid leukemia; 8 T-lineage and 19 B-lineage acute lymphoblastic leukemia [21]
cns	200	42	5	embryonal tumors of the central nervous system (cns): 10 medulloblastomas, 8 primitive neuroectodermal tumors, 10 atypical teratoid/rhabdoid tumors, 10 malignant gliomas, 4 normal cerebellum [36]
novartis	200	103	4	samples from tumor tissues of four different types: 26 breast, 26 prostate, 28 lung, 23 colon [40]

Monti et al., 2003, Machine Learning

Leistungsvergleich von Resampling-Techniken

Daten III: biologische Daten

Name	p	n	C_{biol}	n_i	Comments
iris	4	150	3	50	The first four data sets taken from the UCI repository [6] have been widely used for testing various cluster analysis tools. Each variable of the UCI data sets was normalized to have zero mean and unit variance.
liver	6	345	2	200, 145	
thyroid	5	215	3	150, 35, 30	
wine	13	178	3	59, 71, 48	
fMRI	80	200	4	50	

Blake and Merz, 1998, "UCI repository of machine learning databases."

Möller et al., 2002, NeuroImage

Vergleichende Resampling-Studie

Hauptergebnis

Robustes Ranking:

Fehlerwahrscheinlichkeit

1. Perturbation:

addiere 1%, 5% oder 10% Rauschen 0.05, 0.11, 0.18

2. Subsampling

ziehe zufällig 80% der Datenpunkte 0.21

3. Bootstrapping

ziehe n Datenpunkte mit Zurücklegen 0.33

→ Verlust an Original-Information; ggf. kritisch für die Identifikation kleinerer Klassen bei kleinen Stichproben

Gliederung

Einleitung

- Clustering von Microarray-Daten
- Validierung mittels Resampling

Leistungsvergleich von Resampling-Techniken

- Strategie, Methoden, Daten
- Hauptergebnis

Nächste-Nachbarn-Resampling

- Idee
- Ergebnisse

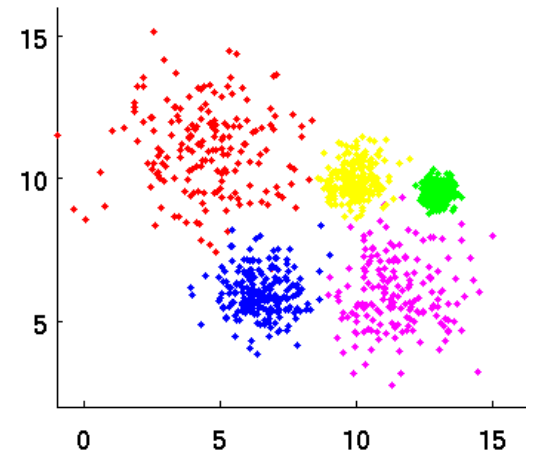
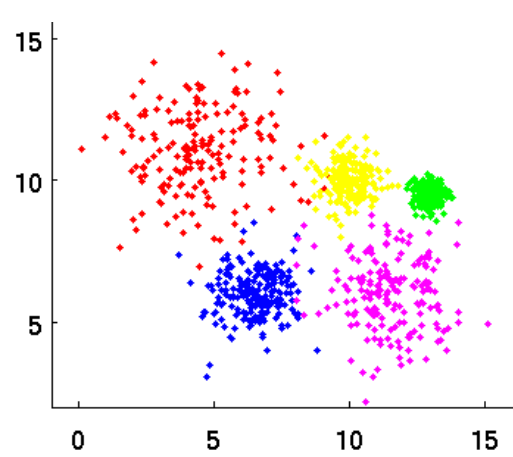
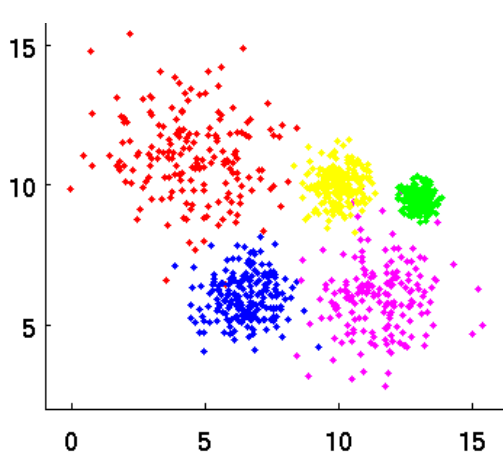
Schlussfolgerungen



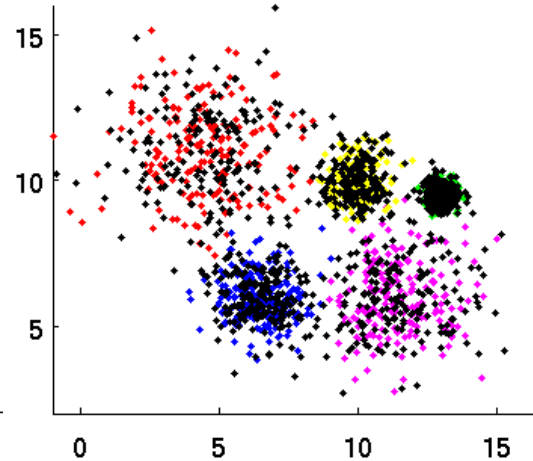
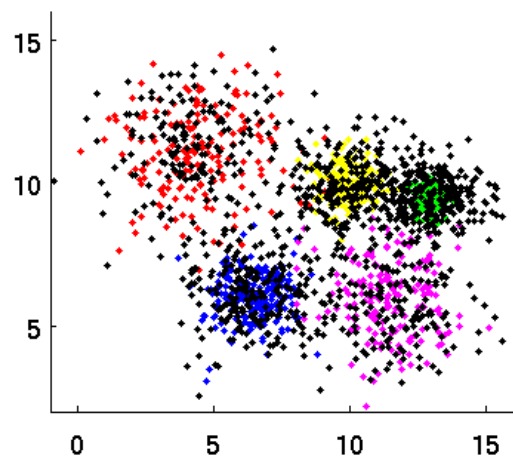
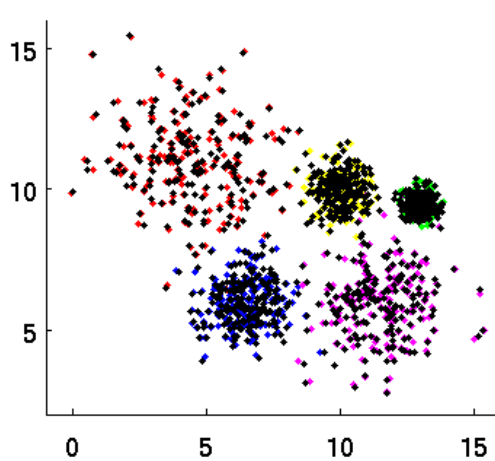
Schätzung lokaler statt globaler Perturbationsparameter

Nächste-Nachbarn-(NN)-basiertes Resampling

Original-
sample



Original-
sample



+ 3% Rauschen

+ 30% Rauschen

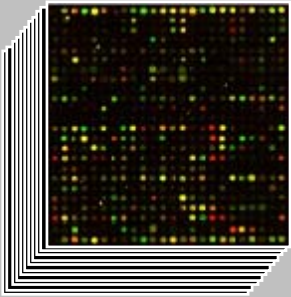
30-NN-resampling

Resample

Stabilitätsanalyse von Resample-Partitionen

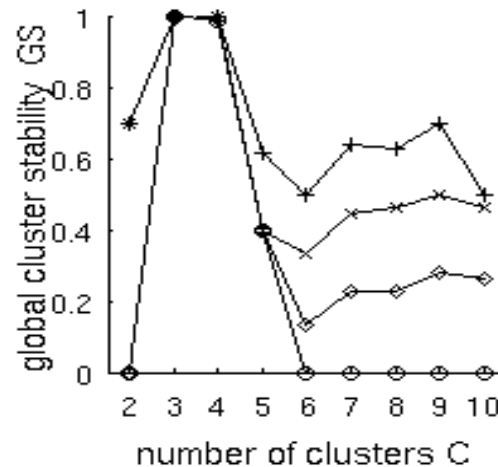
Identifikation von Tumor-Subtypen aus Microarray-Daten

Leukämie



999 Gene
38 Patienten
3 Klassen
• ALL-B
• ALL-T
• AML

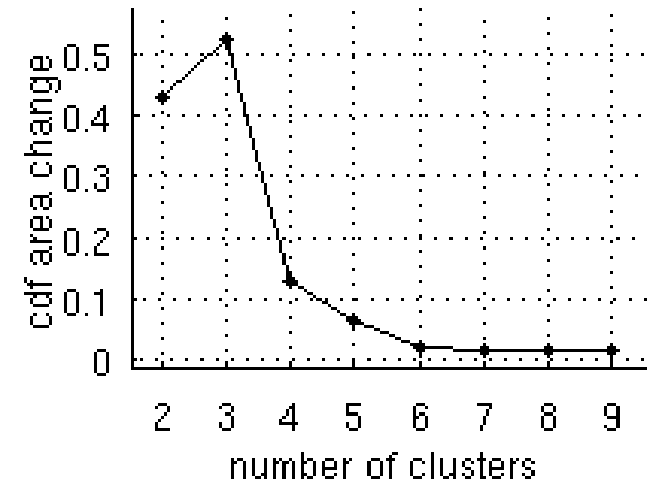
(Golub et al., 1999;
Monti et al., 2003)



Methodenpaket 1

- 10 NN-Resamples
- FCM, mittlere Stabilität ganzer Cluster

Möller U, Radke D,
A cluster validity approach based on
nearest neighbor resampling.
Proc. of the Int. Conf. on Pattern
Recognition (ICPR06), 2006, Hong Kong



Methodenpaket 2

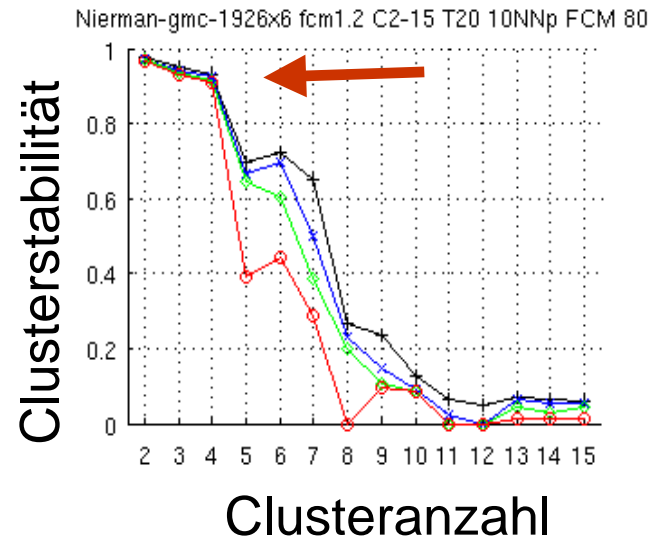
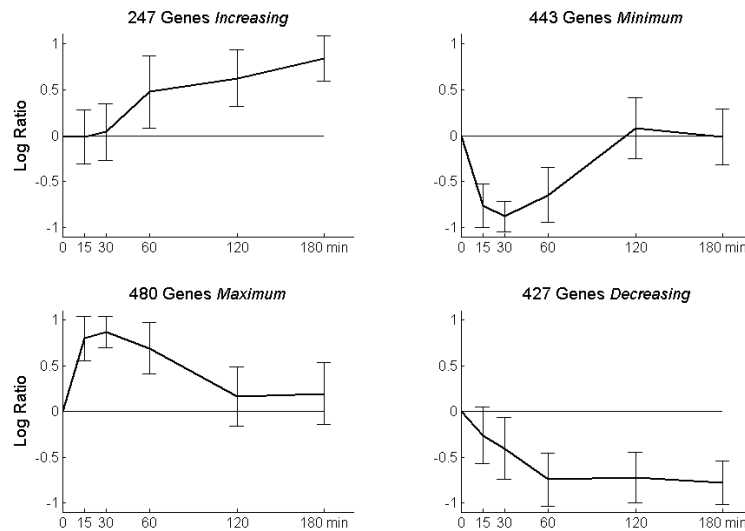
- 500 Subsamples !
- av. link., mittlere Stabilität paarweiser Zuordnungen

Monti et al. Consensus Clustering:
A Resampling-Based Method for Class
Discovery and Visualization of Gene
Expression Microarray Data.
Machine Learning, 2003

Stabilitätsanalyse von Resample-Partitionen

Stabile Gruppen ko-exprimierter Gene

Microarray-Daten einer Temperatur-Verschiebung in *Aspergillus fumigatus* (human-pathogener Pilz) Nierman et al., Nature, 2005



4 klare Muster der Ko-expression

bis zu 4 sehr stabile Cluster

- Guthke R, Kniemeyer O, Albrecht D, Brakhage AA, Möller U: Discovery of gene regulatory networks in *Aspergillus fumigatus*. accepted for Springer Lecture Notes in Bioinformatics, 2006
- Möller, U., Radke, D., Guthke, R.: Stability analysis of gene expression time course clustering used to optimize input data for gene regulatory network reconstruction. Abstracts of the NiSIS/JCB-Workshop 'Top-down Approaches in systems biology - applications', Jena, May 5, 2006, pages 8-10

Schlussfolgerungen

Kontext: kleine Stichprobe ($N < p$), Resampling
(nicht-parametr.), Clustering

- 1) Nutze die **gesamte** Information der Original-Stichprobe X durch eine „1:1“-Abbildung von X auf jedes Resample.
(**bessere** Resample-Qualität, **weniger** Resamples nötig)
- 2) Wenn die Intra-Cluster-Variabilität von Cluster zu Cluster schwanken kann, empfiehlt sich eine **automatische** Schätzung **lokaler** Perturbationsparameter, z.B. mittels **kNN**.
- 3) Durch Suche nach Clustern mit einer intuitiv definierbaren **Mindest-Stabilität** unter Bedingung 1) können Substrukturen in nur teilweise strukturierten Daten besser gefunden werden.