

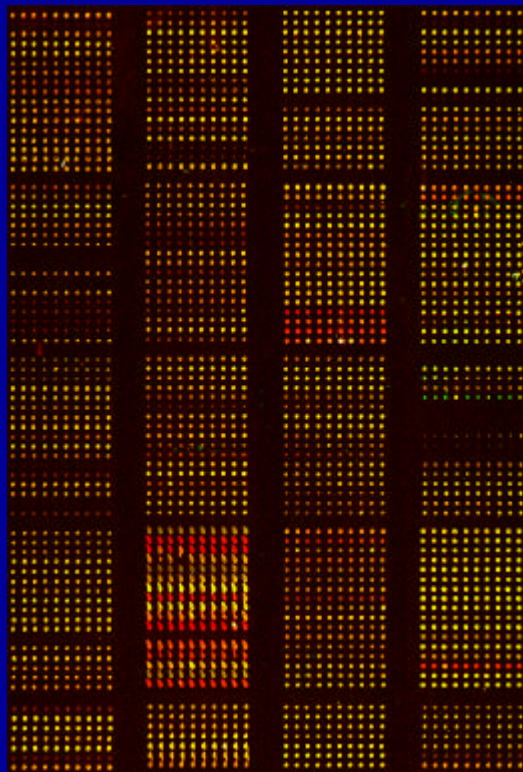
Klassifikations- und Erklärungsmodelle von Tumoren anhand molekulargenetischer Marker

Falk Schubert, Roland Eils
Intelligente Bioinformatiksysteme
Deutsches Krebsforschungszentrum
Heidelberg

Typische Frage in der Bioinformatik

- Zu welcher Tumorart gehört eine Probe?
- Auf welchen Eigenschaften basiert die Klassifikation?
- Welche statistische Sicherheit liegt vor?

Welcher Tumor?



0.43	0.59	0.30	0.00	0.23	0.11	-0.04	0.33
0.21	0.41	-0.13	0.01	0.18	-0.19	missing	0.49
-0.23	0.18	-0.11	-0.10	-0.47	0.18	-0.21	-0.36
0.01	0.35	0.19	0.16	0.03	0.05	-0.08	0.20
0.43	missing	-0.06	0.04	0.14	0.11	0.00	-0.08
0.48	missing	-0.10	0.03	0.20	0.08	-0.09	-0.24
0.21	-0.05	-0.17	0.15	-0.17	0.13	-0.16	-0.40
0.55	0.44	0.22	0.00	0.41	0.00	0.11	0.06
0.09	-0.12	0.24	0.04	-0.04	0.14	-0.13	-0.27
0.06	0.50	-0.13	-0.03	-0.07	0.12	-0.08	-0.03
0.20	0.56	-0.22	0.04	0.07	-0.04	-0.01	0.07
0.09	0.51	0.01	0.01	-0.22	0.37	0.09	0.10
0.11	0.48	0.00	0.02	-0.07	0.26	0.05	0.05
0.17	0.51	0.06	-0.01	-0.06	0.25	0.09	-0.33
0.14	0.41	0.03	-0.03	0.03	0.23	0.11	-0.12
-0.15	0.47	0.02	0.00	-0.36	0.37	-0.17	-0.02
-0.07	0.19	-0.18	0.02	-0.04	0.04	-0.08	-0.19
-0.04	0.40	-0.02	-0.01	0.01	-0.01	-0.12	-0.01
-0.01	missing	0.12	0.00	0.06	0.17	-0.14	-0.20
-0.08	0.40	-0.19	-0.08	-0.02	0.04	-0.23	-0.42
0.04	NIL	-0.37	-0.08	0.26	0.03	-0.27	-0.17
0.09	0.30	0.01	0.11	0.41	0.00	missing	-0.19
0.08	0.24	0.22	0.12	0.05	0.08	-0.23	0.37
-0.04	missing	0.04	0.26	0.22	-0.01	-0.15	-0.07
-0.03	0.23	0.06	0.20	0.20	-0.07	-0.14	0.44
0.03	0.41	-0.20	-0.06	0.27	-0.02	-0.24	0.06
-0.01	0.36	-0.10	0.09	0.06	-0.09	missing	0.01
0.07	0.47	-0.18	0.05	0.10	0.01	-0.16	missing
-0.21	-0.26	0.34	0.11	0.27	0.03	-0.11	-0.23

matrix-CGH

Molekulargenetische Marker

- Genexpressionsprofile
 - Microarrays
- Genomische Profile
 - CGH (comparative genomic hybridisation)
 - matrix CGH
- Hier von zwei Tumorarten
 - dedifferenzierte/pleiomorphe Liposarkome

Workflow

- Datenvorverarbeitung
- Visualisierung / Hypothesengenerierung
- Featureselektion
- Klassifikatorentwurf
- Klassifikatortraining und –validierung
- Erweiterung des Klassifikators um eine Erklärungskomponente

Datenvorverarbeitung

- Fehlende Werte
- Verrauschte Werte
- Inkonsistenzen
- Annotation, Datenformat
- Normalisierung

Visualisierung

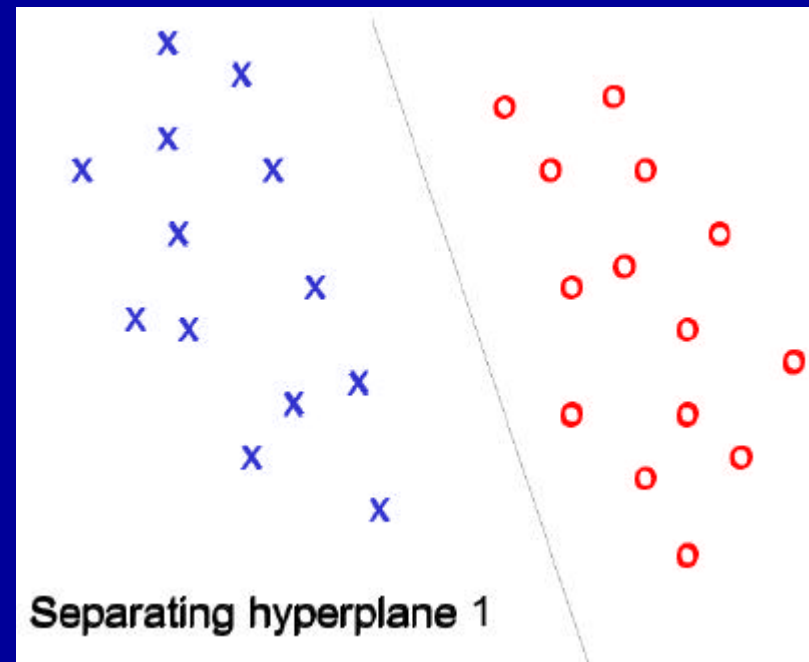
- Beispiel:
 - Daten im 228 dimensionalen Raum
- Projektion / PCA (1./2. Hauptkomponente)
zur Visualisierung nötig

Visualisierung / SOM

- Unüberwachtes Lernen
- Self-organizing map (SOM)
- U-Matrix Darstellung

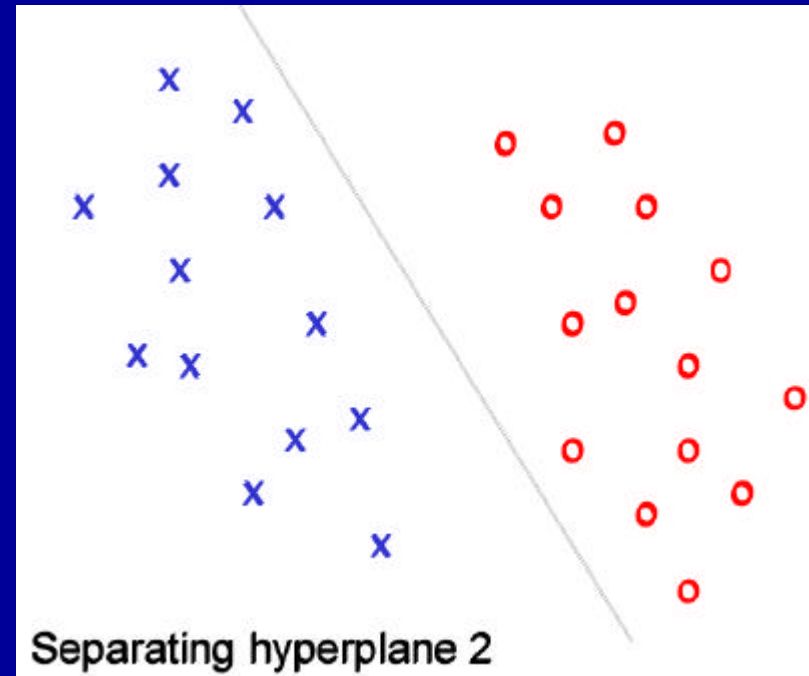
Klassifikation SVM

- Support Vector Machine
- Basierend auf der statist. Lerntheorie (Vapnik)



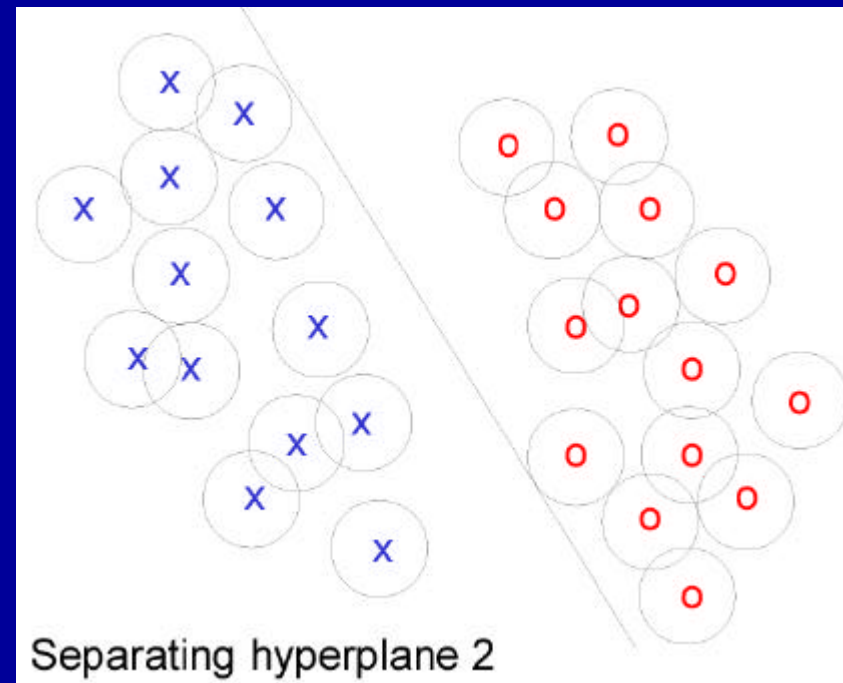
Klassifikation SVM

- Support Vector Machine
- Basierend auf der statist. Lerntheorie (Vapnik)



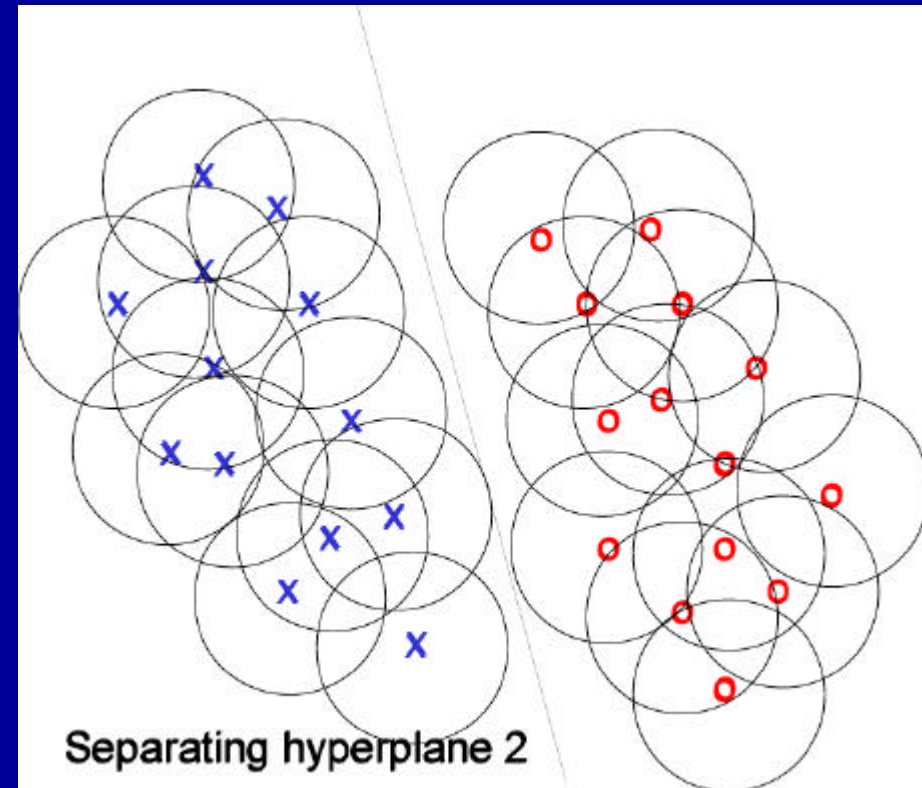
Klassifikation SVM

- Support Vector Machine
- Basierend auf der statist. Lerntheorie (Vapnik)



Klassifikation SVM

- Support Vector Machine
- Basierend auf der statist. Lerntheorie (Vapnik)



Erklärung der Klassifikation

- Ist der Klassifikator kompetent für die Entscheidung?
- Welche Merkmale (Gene) waren für die Klassifikation entscheidend?
- Wie sicher ist die Klassifikation?

Herausforderungen

- Adaption von Algorithmen und Metriken an biologische Messgrößen
 - Wahl des Kernels oder einer Metrik
- Schätzen von Fehlermodellen für molekularbiologische Untersuchungen
 - cDNA microarrays Normalisierung
- Problem der Dimensionalität (curse of dimensionality, empty space phenomenon)

Ziel

- Wissensbasierte Entscheidungsunterstützung in der Klinik
- Unterstützung des Kliniklers bei Interpretation von genomischen Profilen und Genexpressionsprofilen
- Individualisierung der medizinischen Behandlung (Single Nucleotide Polymorphisms)

Danke!

Referenz

Microarray based Copy Number and
Expression Profiling in Dedifferentiated and Pleomorphic Liposarcoma

Björn Fritz, Falk Schubert, Gunnar Wrobel, Carsten Schwaenen, Swen
Wessendorf, Michelle Nessler, Christian Korz, Ralf J. Rieker, Kate
Montgomery, Gunhild Mechttersheimer, Roland Eils, Stefan Joos, and Peter
Lichter

Cancer Research Juli 2002