Improved Gene Selection for Classification of Microarrays

Jochen Jäger — MPI Berlin
Rimli Sengupta — IIT Kharagpur
Walter L. Ruzzo — University of Washington

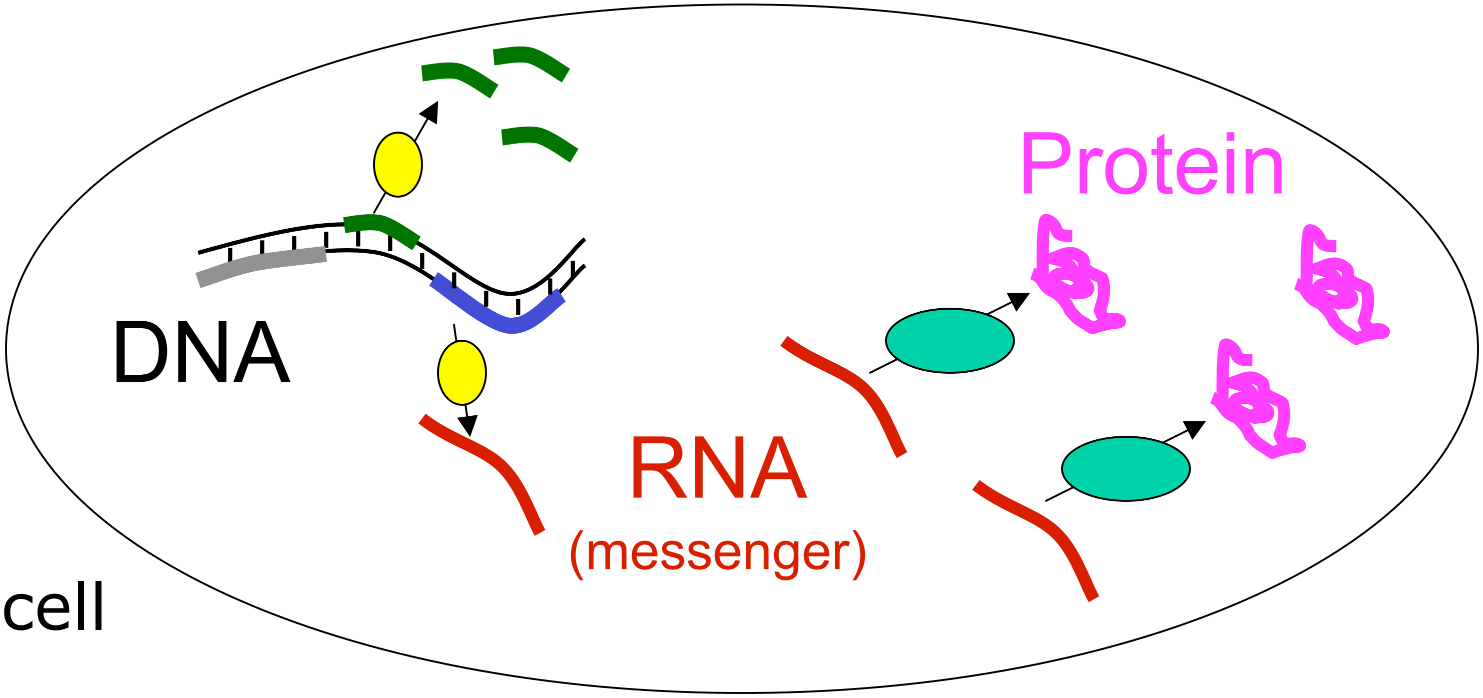
UW CSE Computational Biology Group

Overview

- Gene Expression Microarrays
- Classification and Feature Selection
- One Problem & Three Approaches
- Results
- Summary and Conclusions

Gene Expression: The "Central Dogma"

DNA → RNA → Protein



Gene Expression

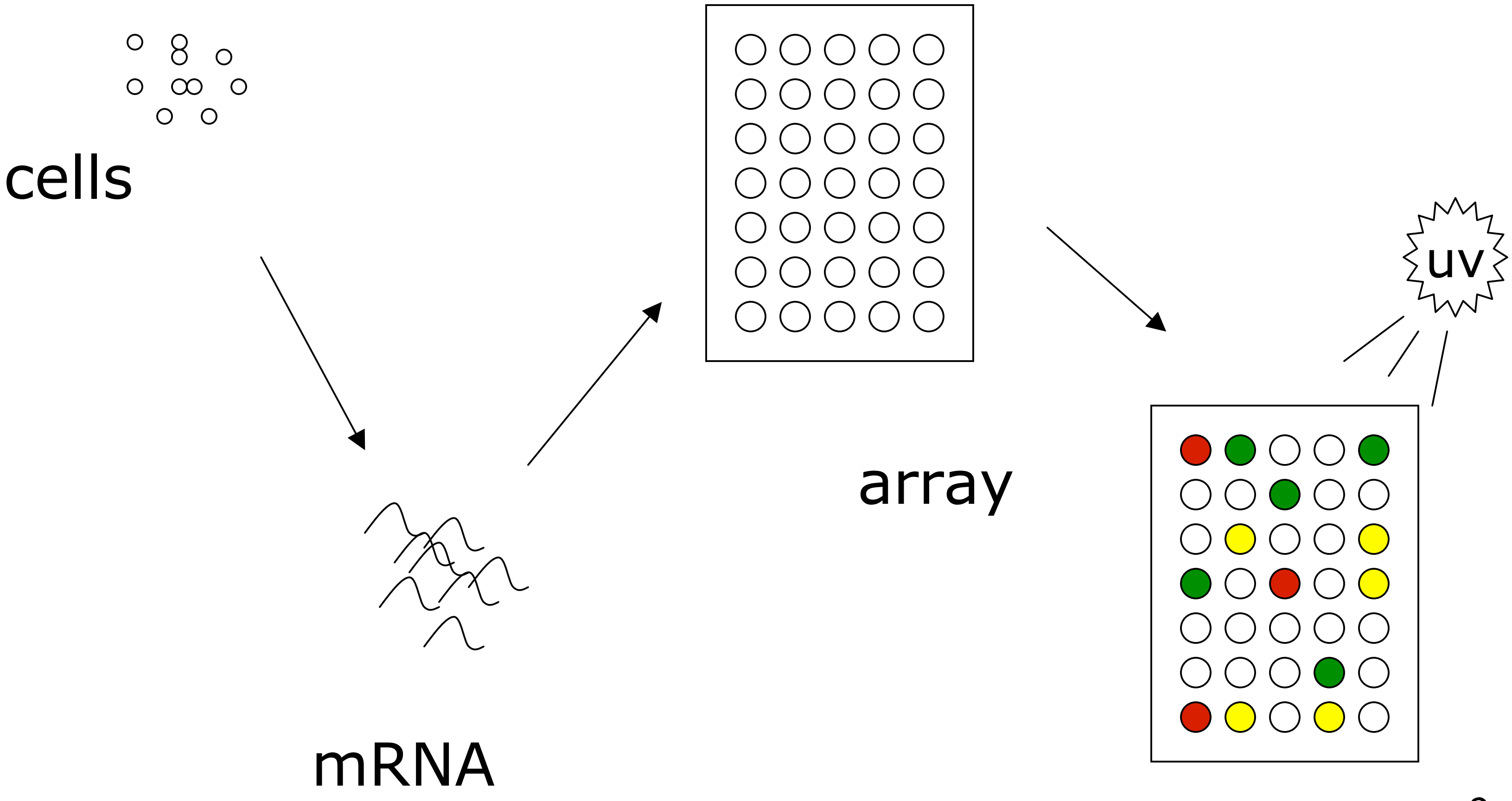
- Proteins do most of the work
- They’re dynamically created/destroyed
- So are their mRNA blueprints
- Different mRNAs expressed at different times/places
- Knowing mRNA “expression levels” tells a lot about the state of the cell

Expression Microarrays

- Thousands to hundreds of thousands of spots per square inch
- Each holds millions of copies of a DNA sequence from one gene

- Take mRNA from cells, put it on array
- See where it sticks – mRNA from gene x should stick to spot x

An Expression Array Experiment

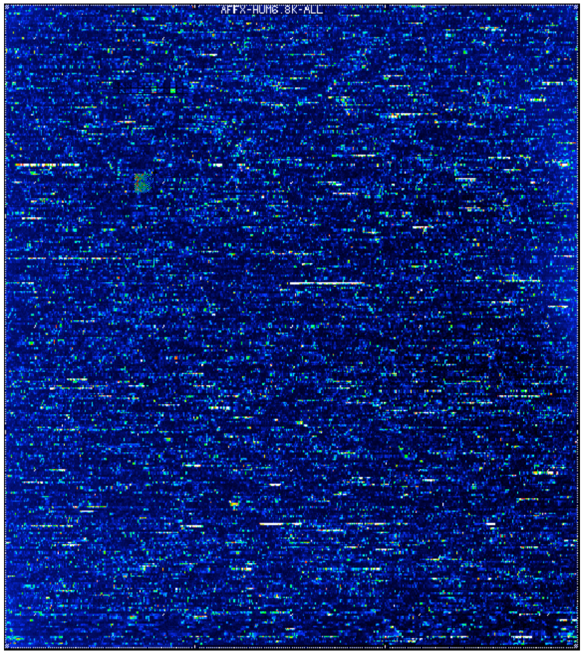


Overview

- Gene Expression Microarrays
- → Classification and Feature Selection
- One Problem & Three Approaches
- Results
- Summary and Conclusions

An Example Application

- 72 leukemia patients
  - 47 ALL
  - 25 AML
- 1 chip per patient
- 7132 human genes per chip



Golub, et al., Science 286:531-537 (1999).

Key Issue:
What's Different?

- What genes are behaving differently between ALL & AML
  (or other disease/normal states)?
- Potential uses:
  - Diagnosis
  - Prognosis
  - Insight into underlying biology/biologies
  - Treatment

A Classification Problem

- Given an array from a new patient: is it ALL or AML?

- Many possible approaches:
  LDA, logistic regression, NN, SVM, …
- Problems:
  - Noise
  - Dimensionality

Feature Selection

- Base the classification on only a subset of the genes
  - Reduce dimensionality – for convenience
  - Drop noisy/irrelevant genes – for accuracy
- Perhaps a very small subset
  - For cost
  - For workload
  - For biological insight

Simple Feature Selection

- Rank genes based on their individual predictive ability, e.g. by t-test or other statistic
- Keep only the top k genes
  + simple, easy, commonly used
  – often highly correlated, so little extra info

An Example

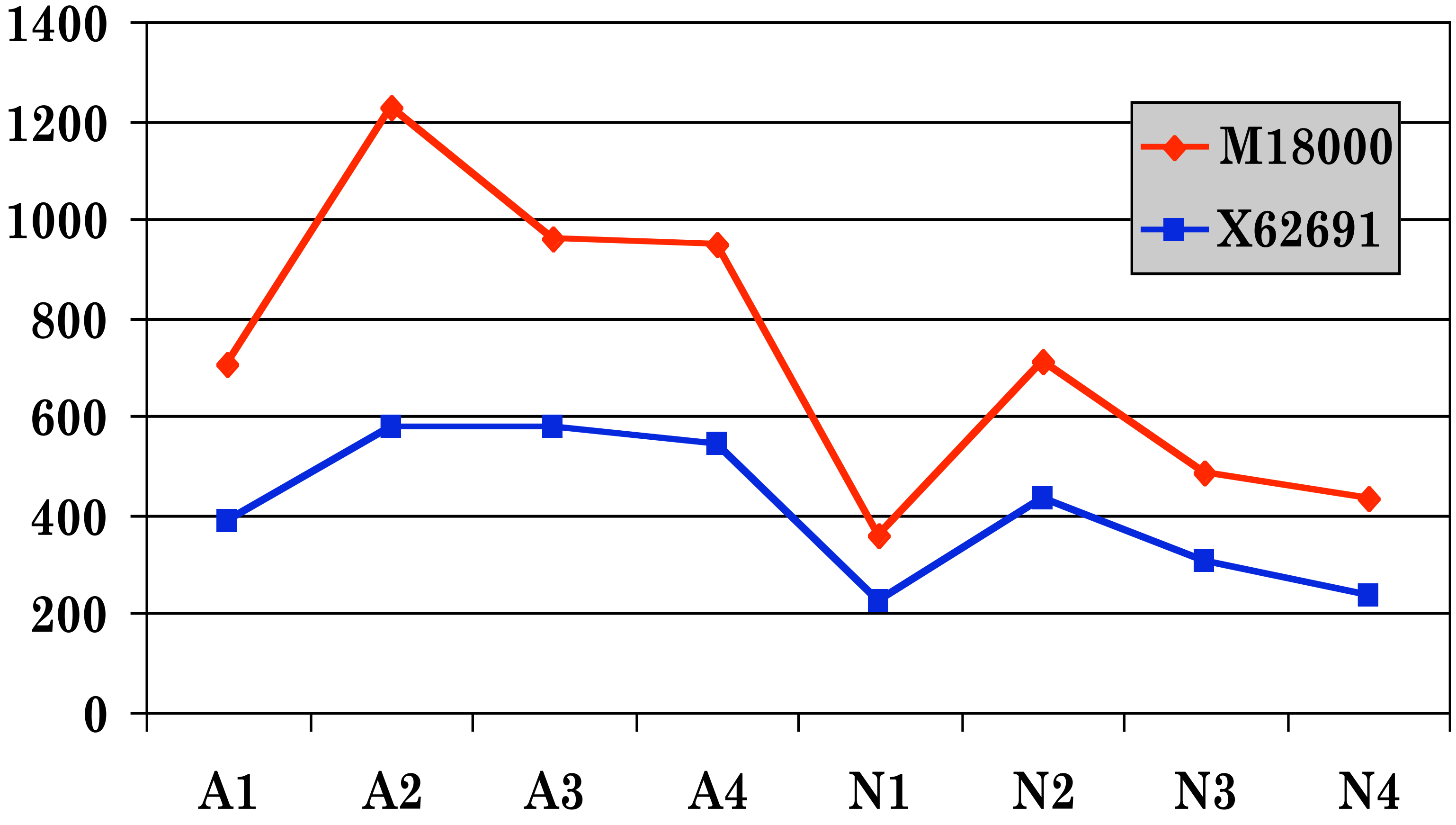
Accession Number	Adenoma				Normal□				t-test p-value
	1	2	3	4	1	2	3	4	
M18000	705.41	1227.27	959.35	951.56	359.83	711.08	485.33	431.19	0.014
X62691	387.91	577.57	578.45	546.54	227.26	436.65	306.94	239.33	0.016
M82962	91.85	16.27	12.61	61.62	187.44	76.90	181.38	186.53	0.017
U37426	0.47	7.05	6.30	3.40	-3.88	1.58	-2.99	-2.91	0.018
HG2564	2.33	0.54	1.58	3.82	-2.91	-2.11	1.00	-2.91	0.019
Z50853	35.43	26.03	51.49	41.22	27.68	15.80	12.46	15.99	0.022
M32373	-48.02	-28.20	-64.62	-56.95	-15.05	-16.86	-7.97	-34.88	0.022

An Example (cont.)

□

	M18000	X62691	M82962	U37426	HG2564	Z50853	M32373
M18000	1.000		□	□	□	□	□
X62691	0.961	1.000		□	□	□	□
M82962	-0.944	-0.971	1.000		□	□	□
U37426	0.973	0.975	-0.983	1.000		□	□
HG2564	0.592	0.653	-0.553	0.529	1.000		□
Z50853	0.514	0.616	-0.633	0.597	0.614	1.000	
M32373	-0.509	-0.590	0.602	-0.580	-0.619	-0.874	1.000

Example


1400
1200
1000
800
600
400
200
0
A1
A2
A3
A4
N1
N2
N3
N4
M18000
X62691

Problem with the simple solution

- Each gene independently scored
- Top k ranking genes might be very similar and therefore no additional information gain
- Reason: genes in similar pathways probably all have very similar score
- What happens if several pathways involved in perturbation but one has main influence
- Possible to describe this pathway with fewer genes

Overview

- Gene Expression Microarrays
- Classification and Feature Selection
- One Problem & Three Approaches
- Results
- Summary and Conclusions

Three Approaches

A: A greedy algorithm picks low p-values and not too high correlation

B: Cluster genes; pick representatives from each cluster

C: Like B, but “mask out” (omit) clusters having poor p-values

Goal of all 3: broader representation of informative genes & pathways

A: “Correlation”

- First gene picked is the one with best p-value
- kth gene picked is the one with best p-value among genes having correlation less than threshold □to previous k-1

B: “Clustering”

- Cluster genes into g groups
- From each cluster, select one or more genes, choosing those with lowest p-values
- Take more from clusters with broad dispersion, fewer from tight clusters (which are likely to be highly correlated)

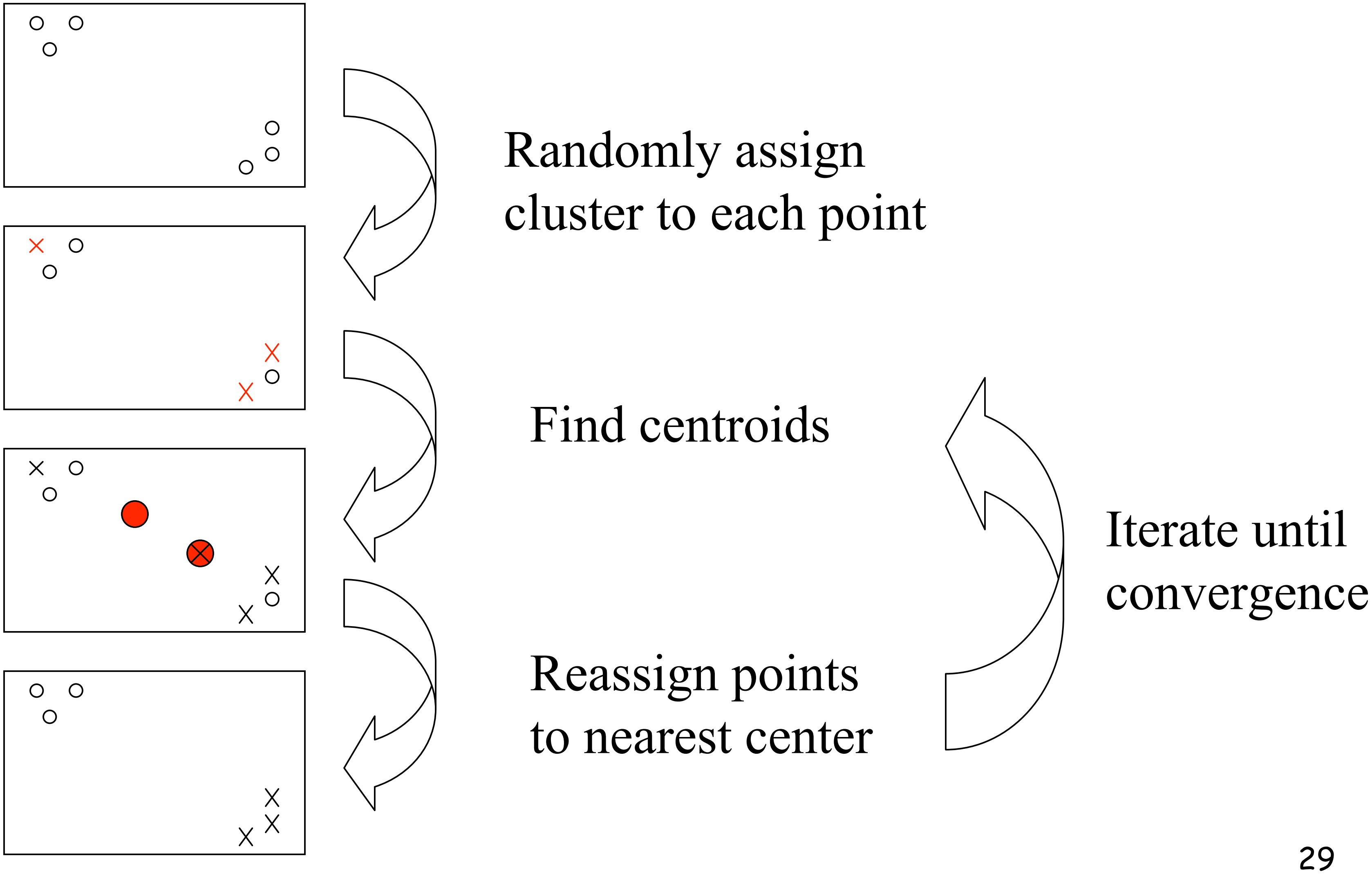
C: “Masked out Clustering”

- Just like B, but don’t take any genes from clusters whose average p-value is poor (> 0.2).

Clustering Algorithms

- K-means
- “Fuzzy” k-means

Hard clustering – k-means



Randomly assign cluster to each point

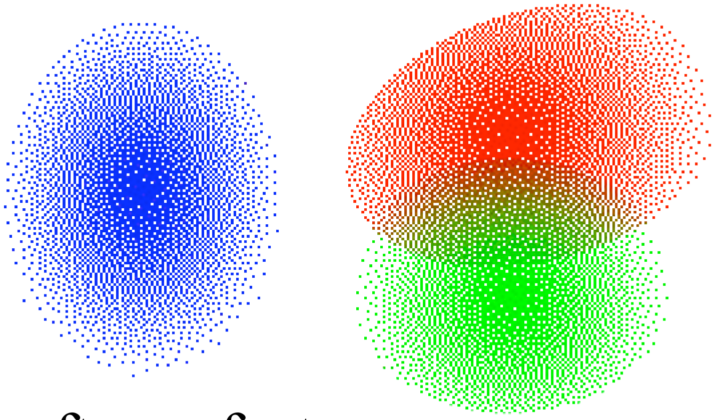
Find centroids

Reassign points to nearest center

Iterate until convergence

Soft - Fuzzy Clustering

instead of hard assignment,
probability for each cluster



Very similar to k-means but fuzzy softness factor m (between 1 and infinity) determines how hard the assignment has to be

Fuzzy examples

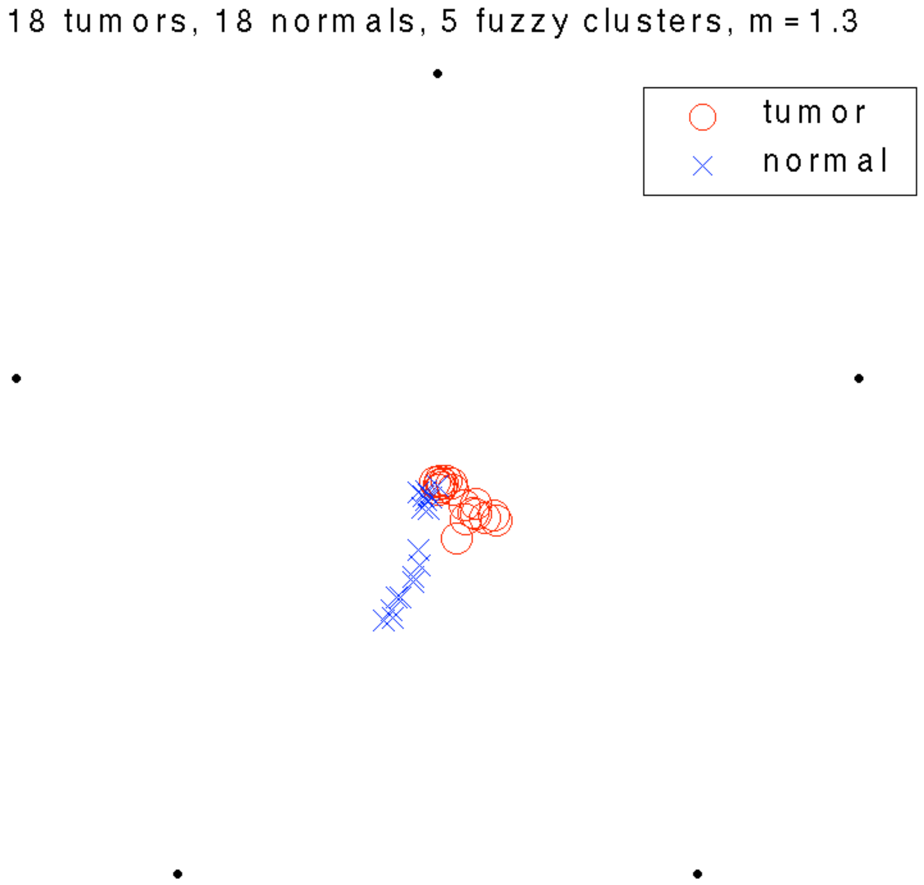
Nottermans carcinoma dataset:

18 colon adenocarcinoma and 18 normal tissues

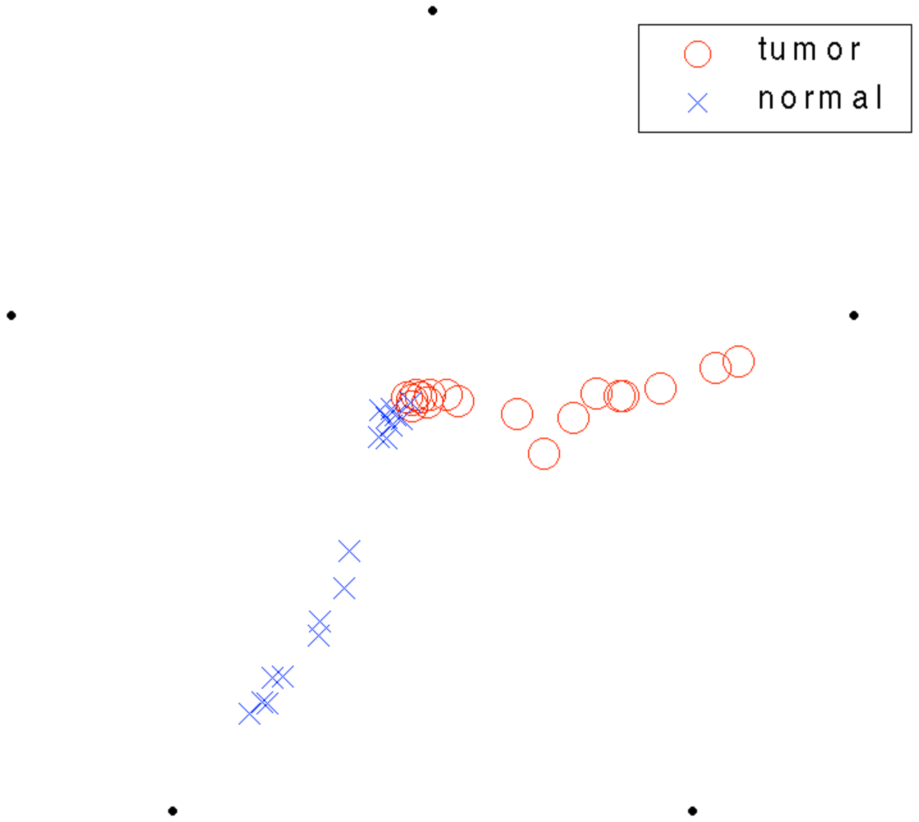
data from 7457 genes and ESTs

cluster all 36 tissues

Fuzzy softness 1.3

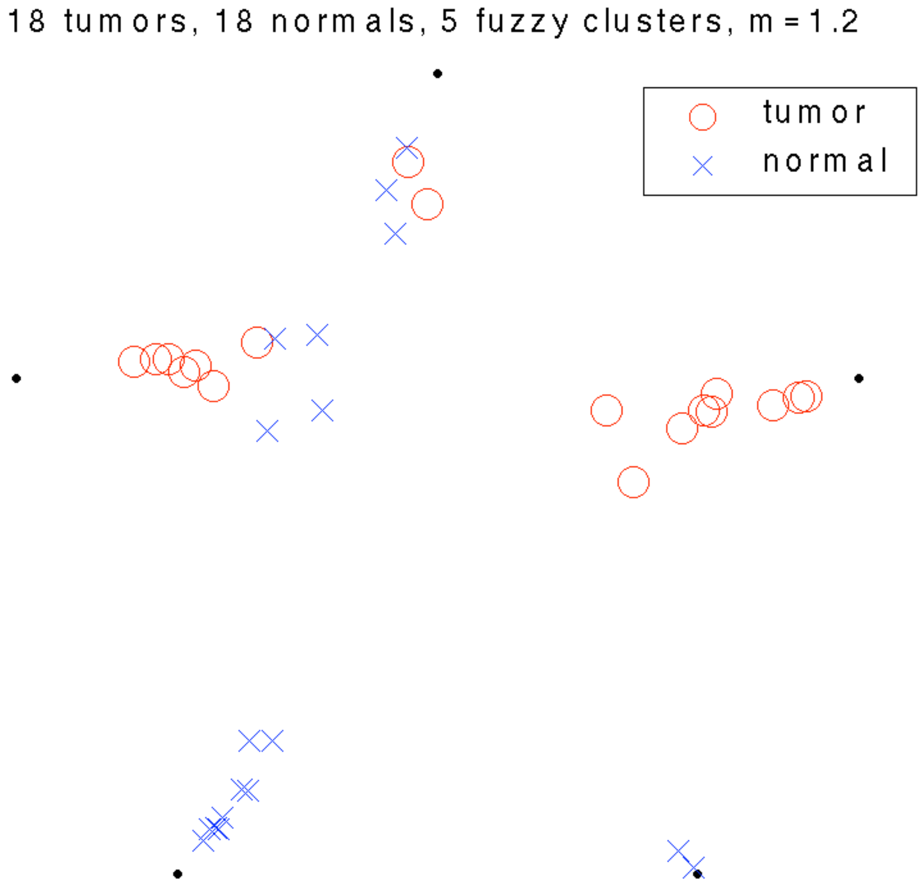


Fuzzy softness 1.25

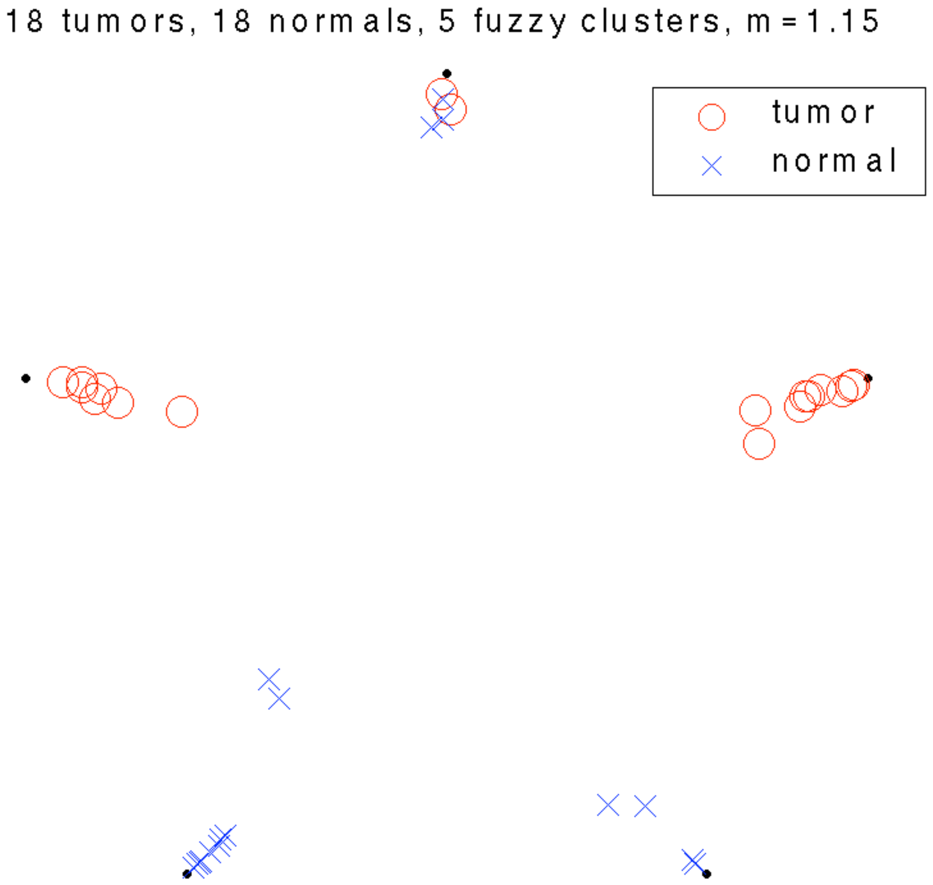


18 tumors, 18 normals, 5 fuzzy clusters, m = 1.25

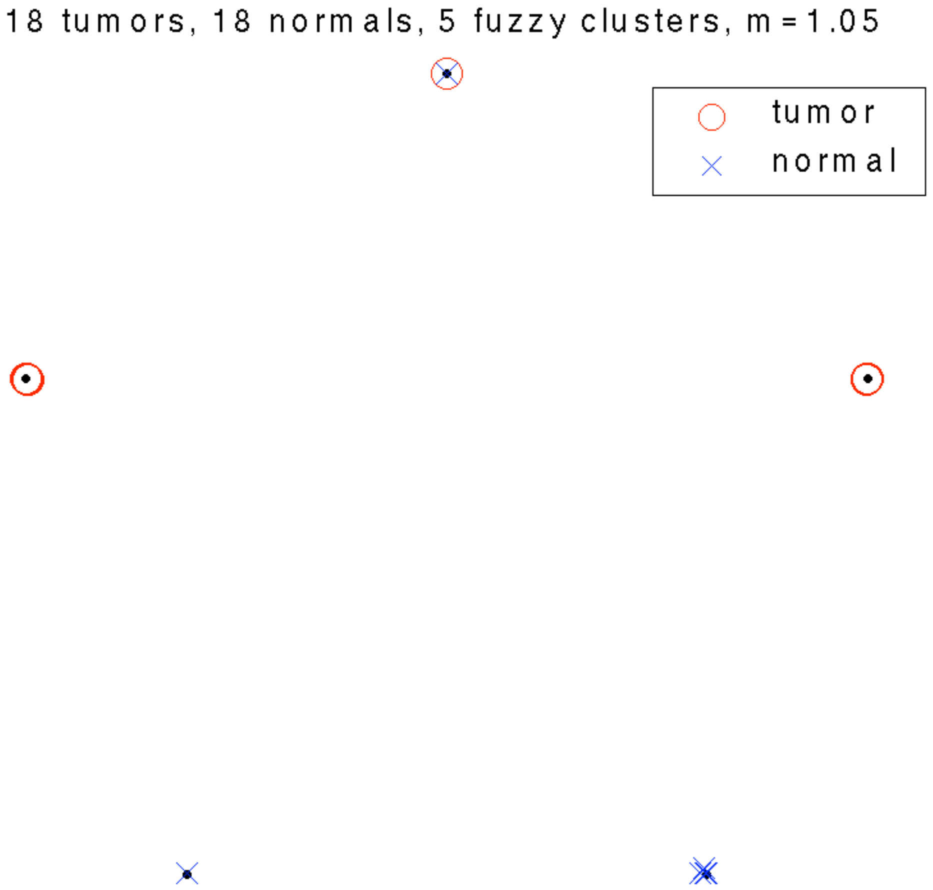
Fuzzy softness 1.2



Fuzzy softness 1.15



Fuzzy softness 1.05



Selecting genes from clusters

- Two way filter: exclude redundant genes, select informative genes
- Get as many pathways as possible
- Consider cluster size and quality as well as discriminative power

How many genes per cluster?

- Constraints:
  - minimum one gene per cluster
  - maximum as many as possible
- Take genes proportionally to cluster quality and size of cluster
- Take more genes from bad clusters
- Smaller quality value indicates tighter cluster

- Quality for k-means: sum of intra cluster distance

Which genes to pick?

- Choices:
  - Genes closest to center
  - Genes farthest away
  - Sample according to probability function
  - → Genes with best discriminative power

Overview

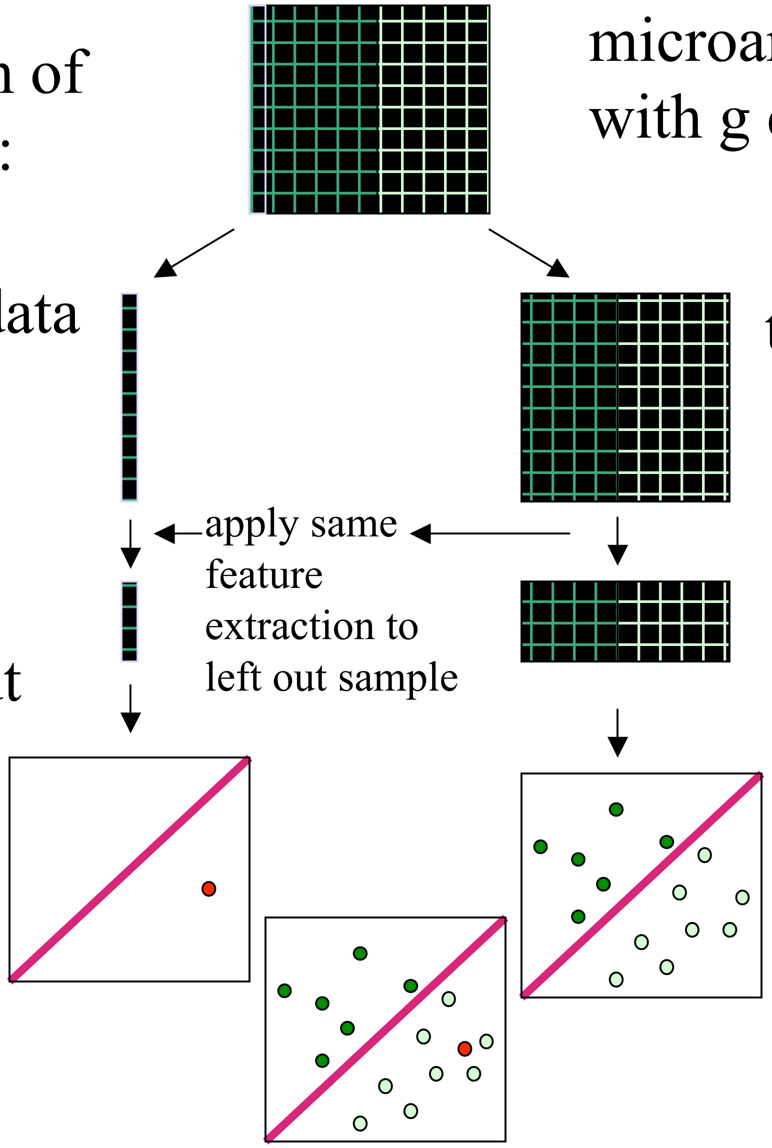
- Gene Expression Microarrays
- Classification and Feature Selection
- One Problem & Three Approaches
- → Results
- Summary and Conclusions

Experimental setup

- Datasets:
  - Golub, et al.: Leukemia (47 ALL, 25 AML)
  - _ Alon, et al.: Colon (40 tumor and 22 normal colon adenocarcinoma tissue samples)
  - Notterman, et al.: Carcinoma and Adenoma (18 adenocarcinoma, 4 adenomas and paired normal tissue)
- Experimental setup:
  - calculate LOOCV using SVM on feature subsets
  - do this for feature size 10-100 (in steps of 10) and 1-30 clusters

Comparison Evaluation

Repeat for each of the n examples: leave out one sample



microarray data: n examples with g expression levels each

test data

train data

extract features

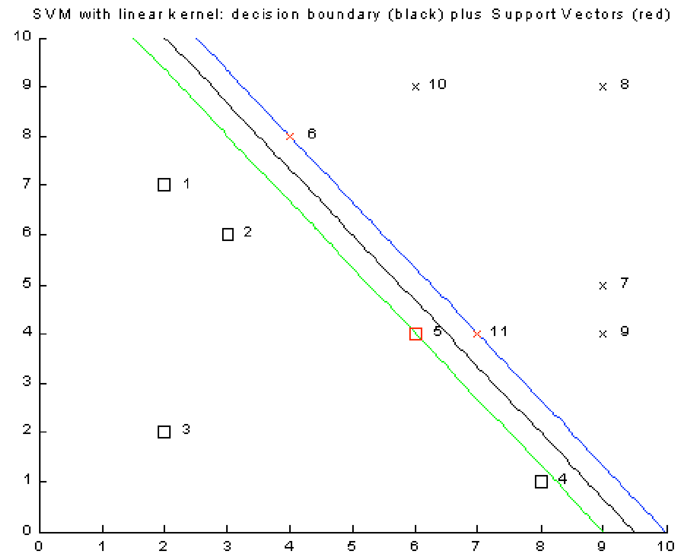
apply same feature extraction to left out sample

classify held-out sample

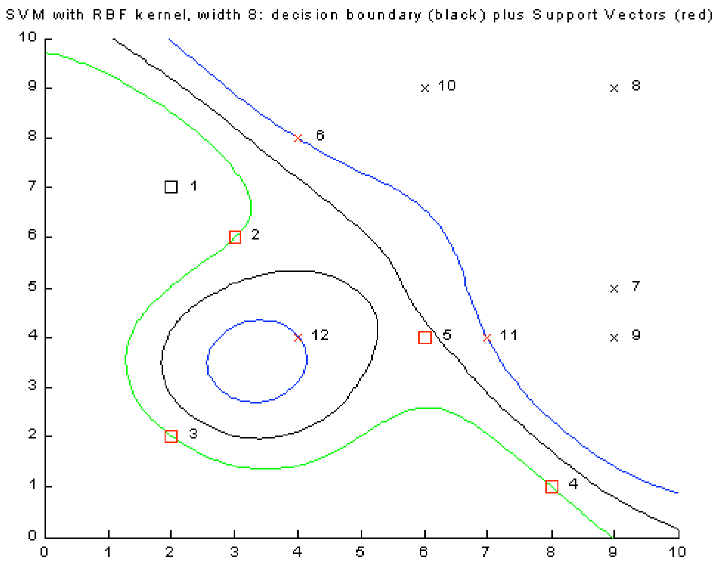
train learner

Support Vector Machines

- Find separating hyperplane with maximal distance to closest training example



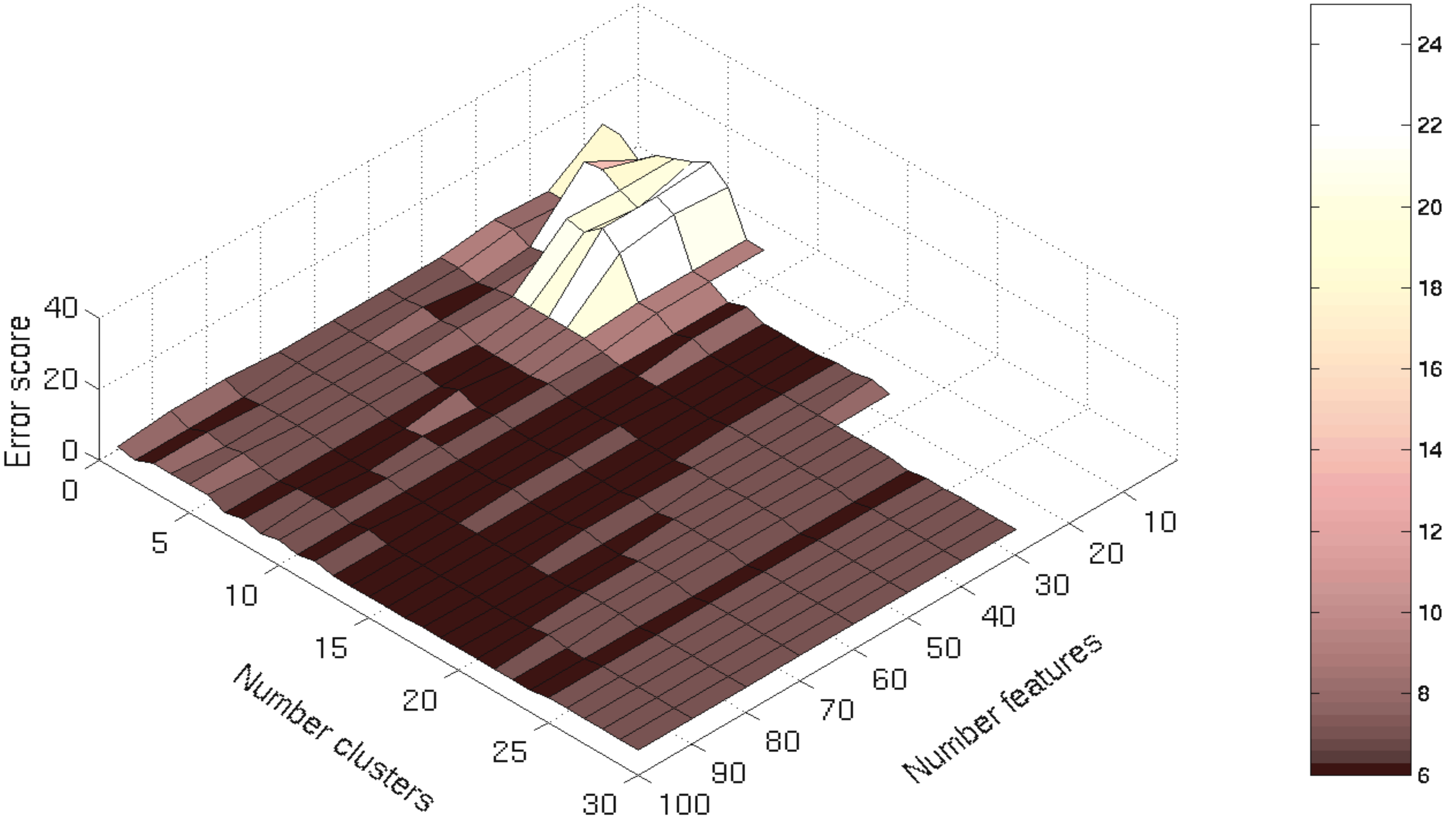
SVM with linear kernel: decision boundary (black) plus Support Vectors (red)



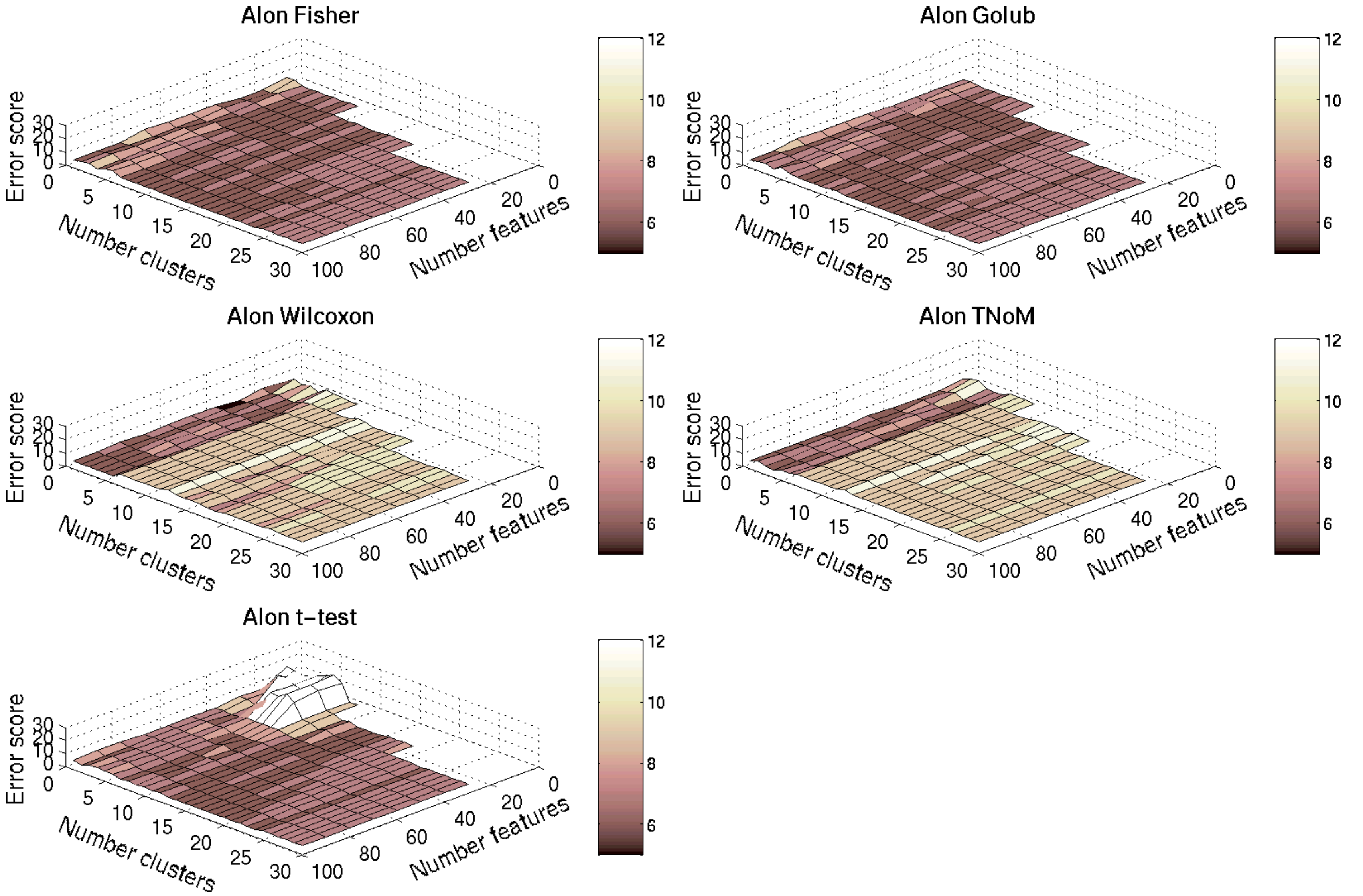
SVM with RBF kernel, width 8: decision boundary (black) plus Support Vectors (red)

  - avoids overfitting
  - can handle higher order interactions and noise using kernel functions and soft margin

Results: Alon,Fuzzy,t-test



Alon, Fuzzy, Other Stats


Alon Fisher

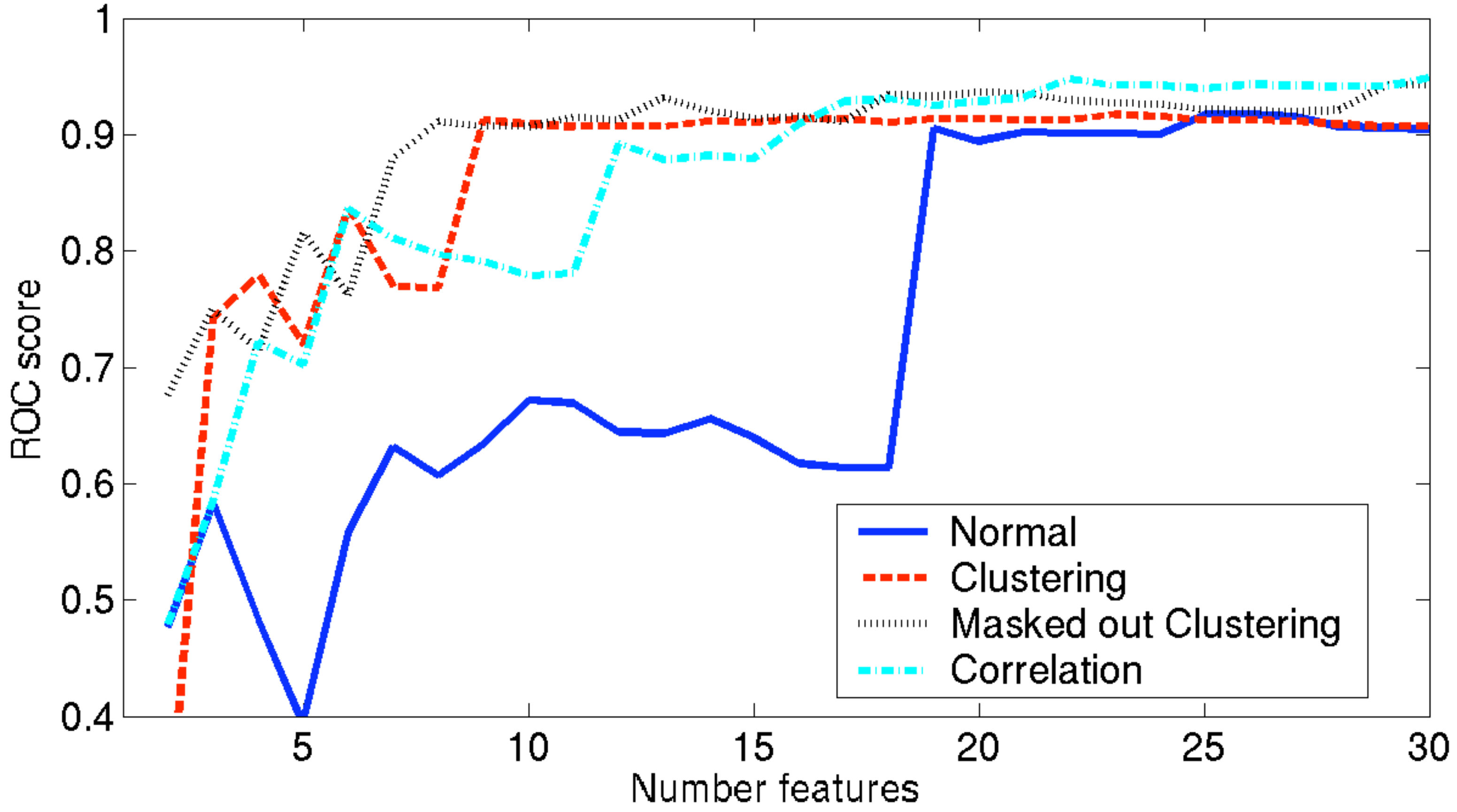
Alon Golub

Alon Wilcoxon

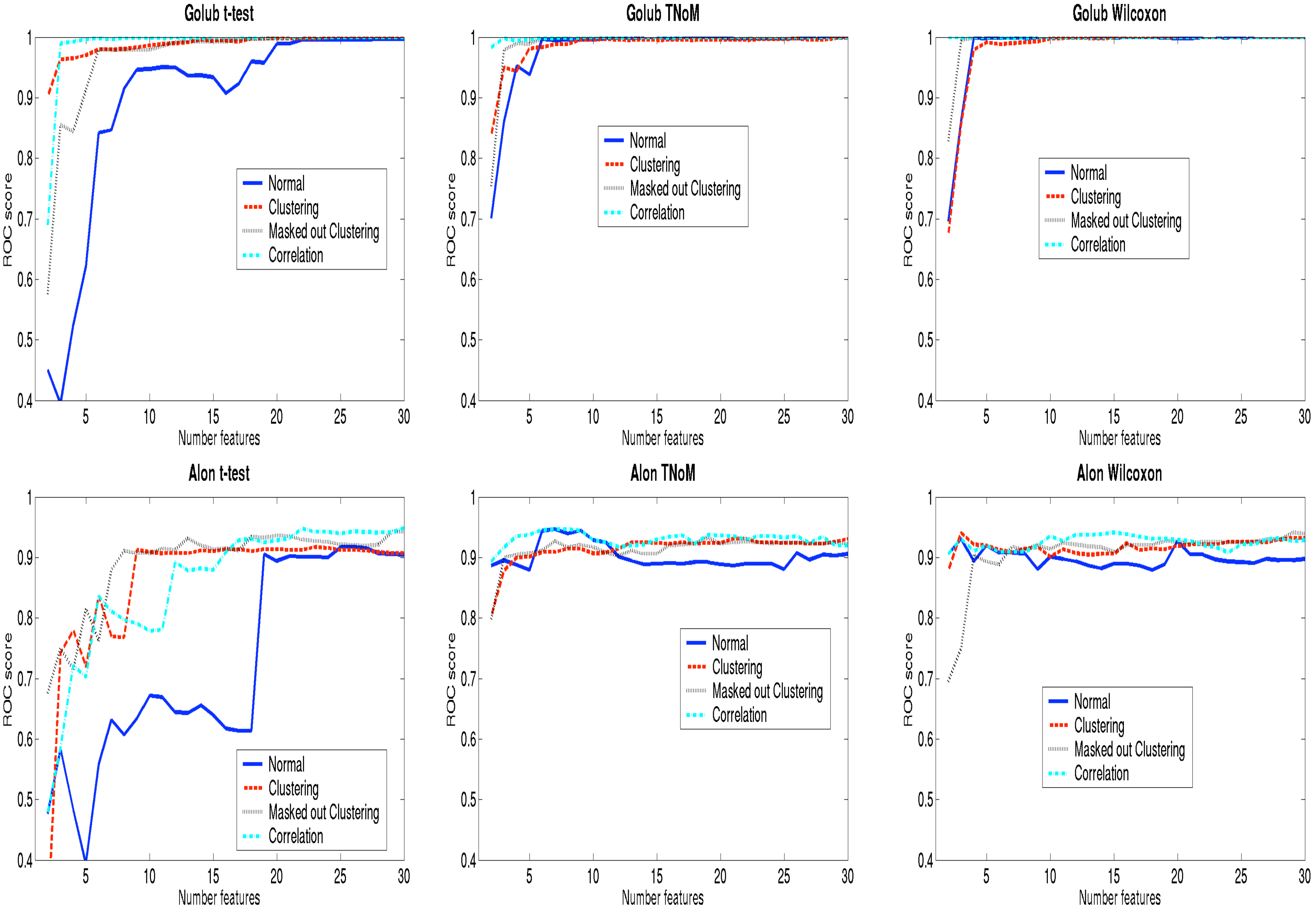
Alon TNoM

Alon t−test

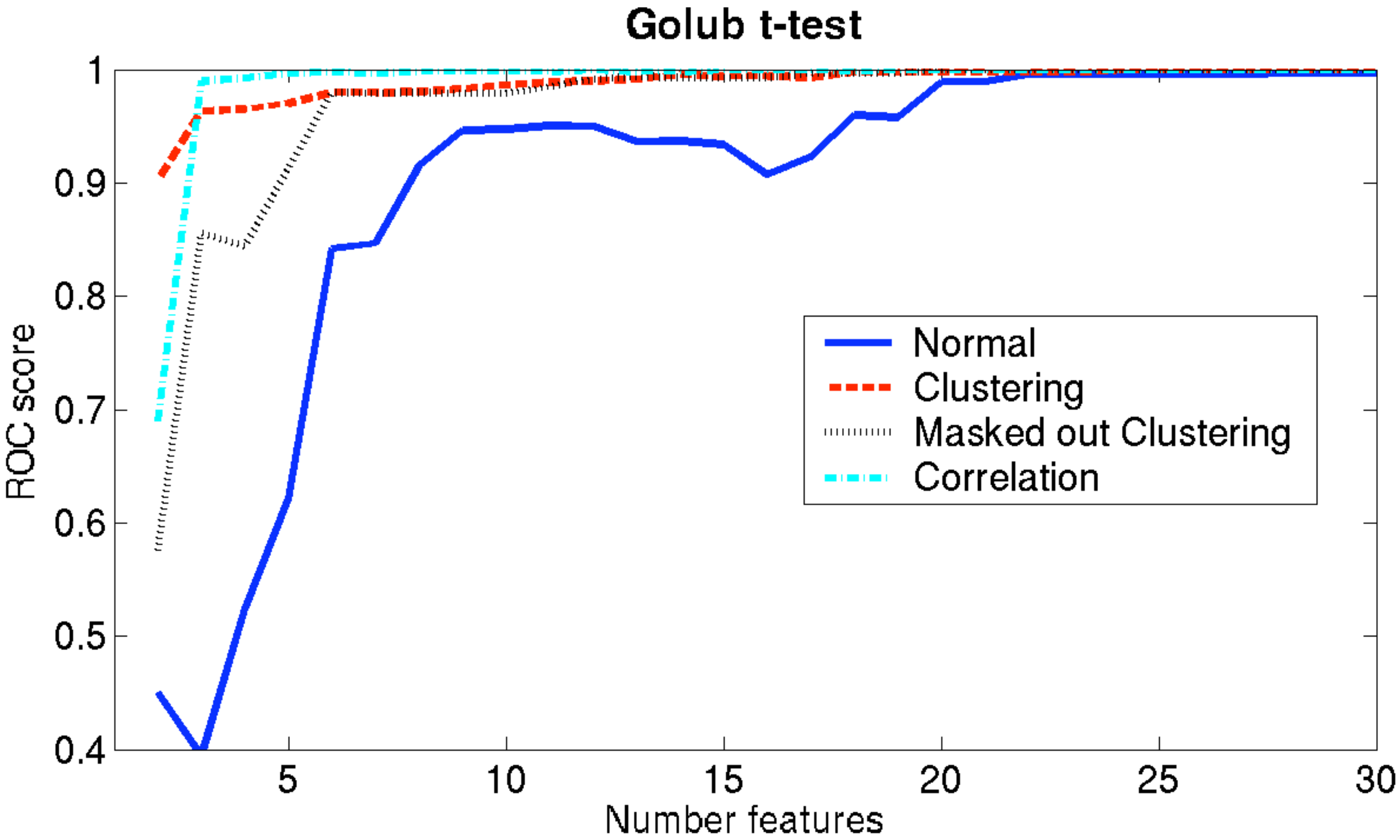
ROC Scores: Alon, t-test



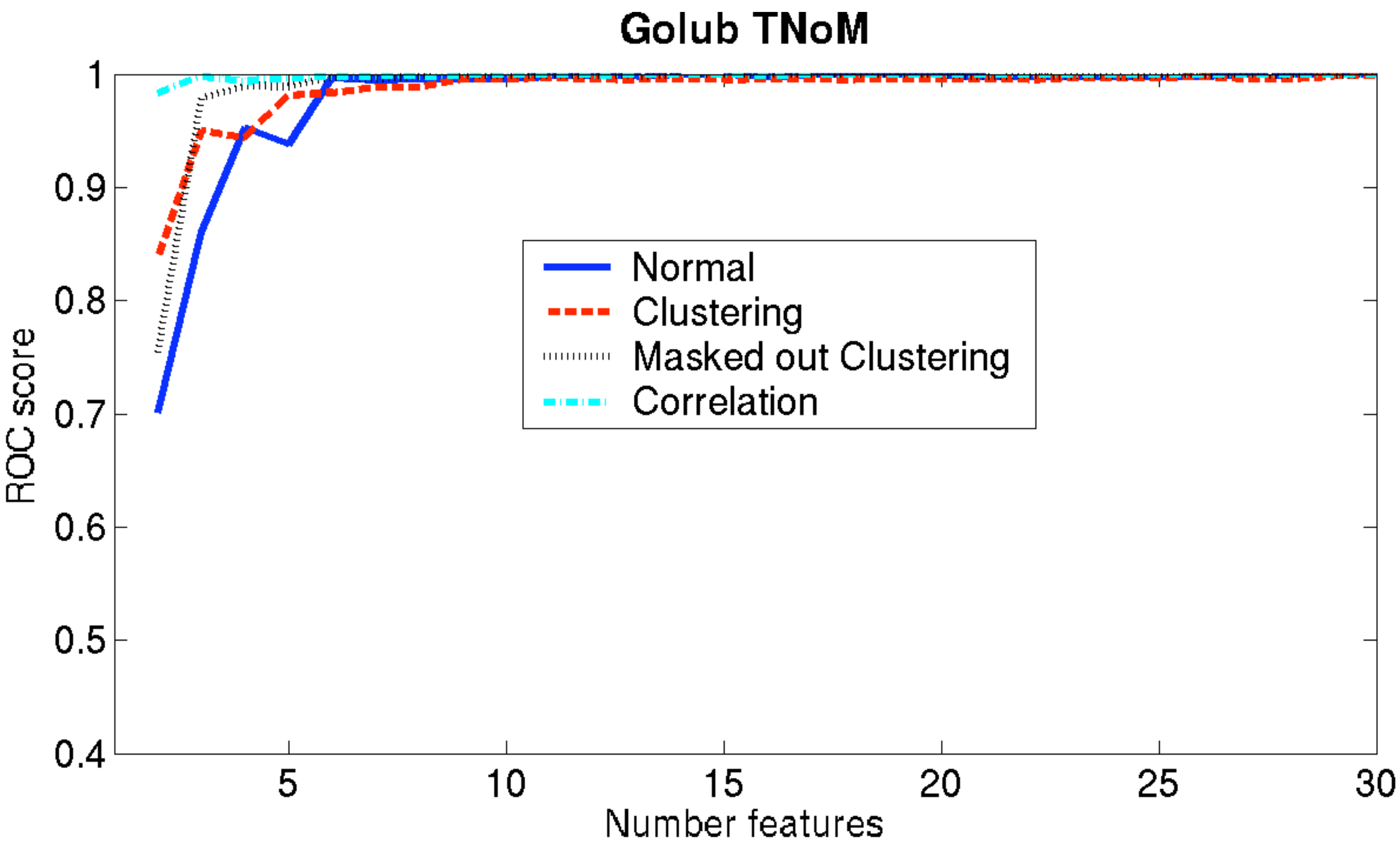
More ROC Scores



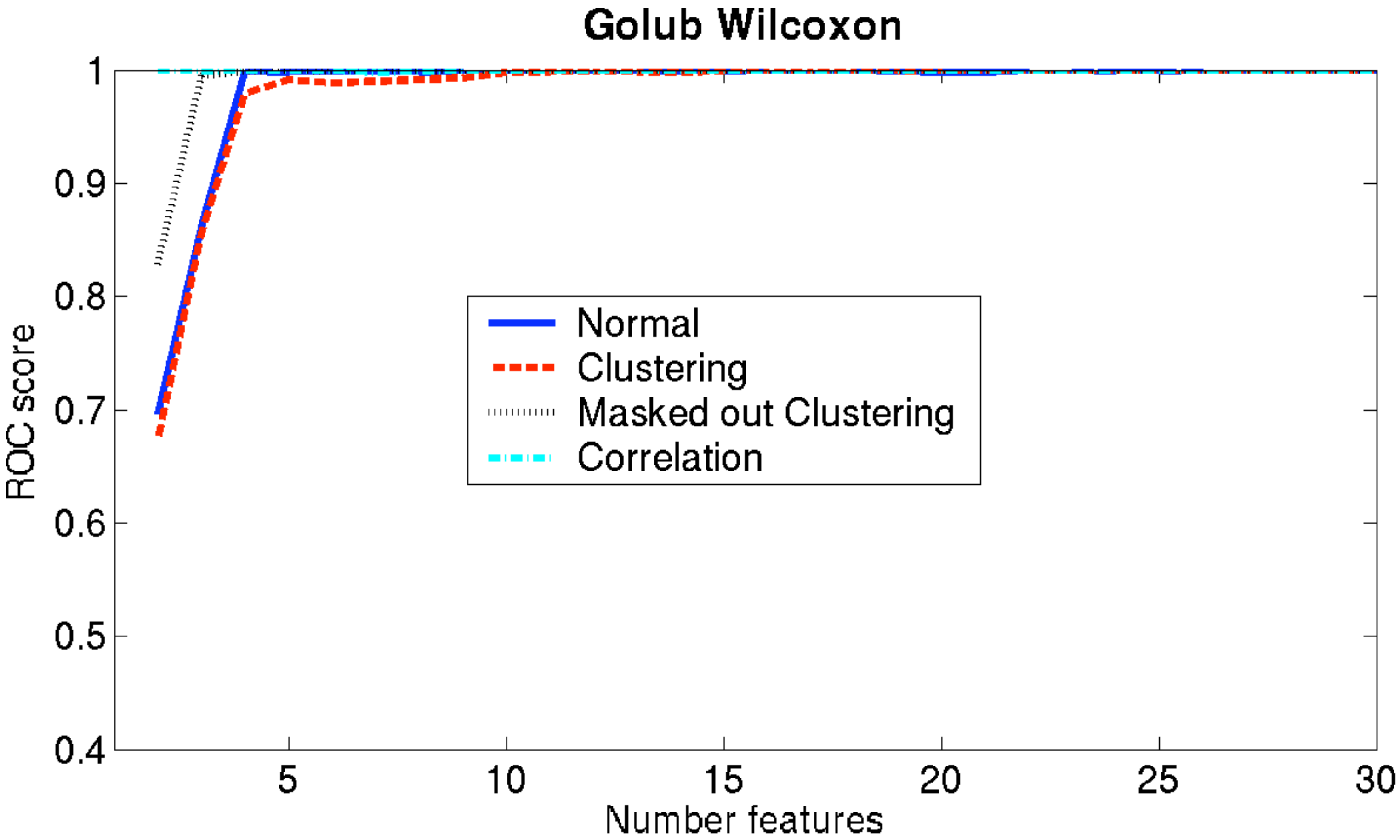
More ROC Scores



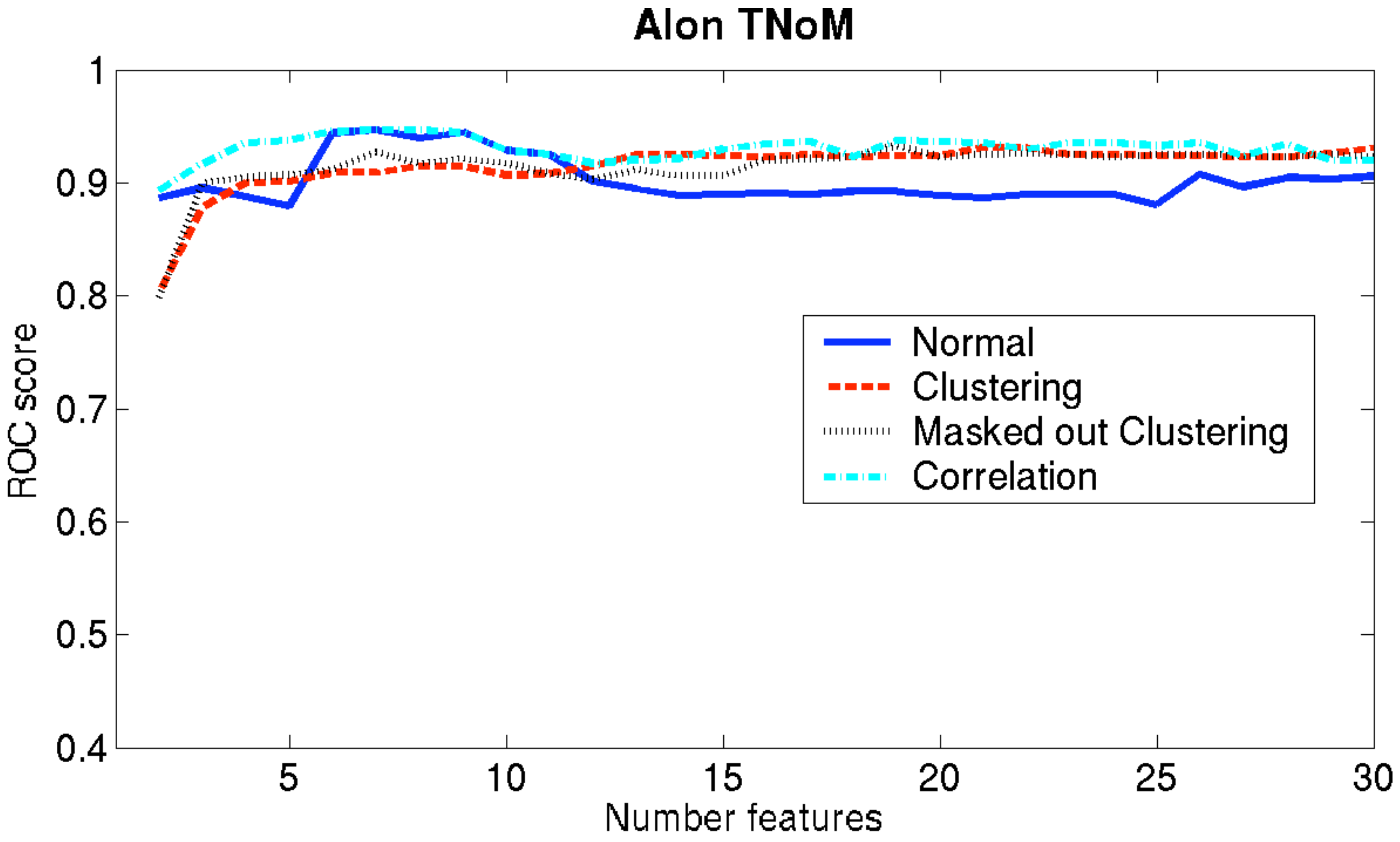
More ROC Scores



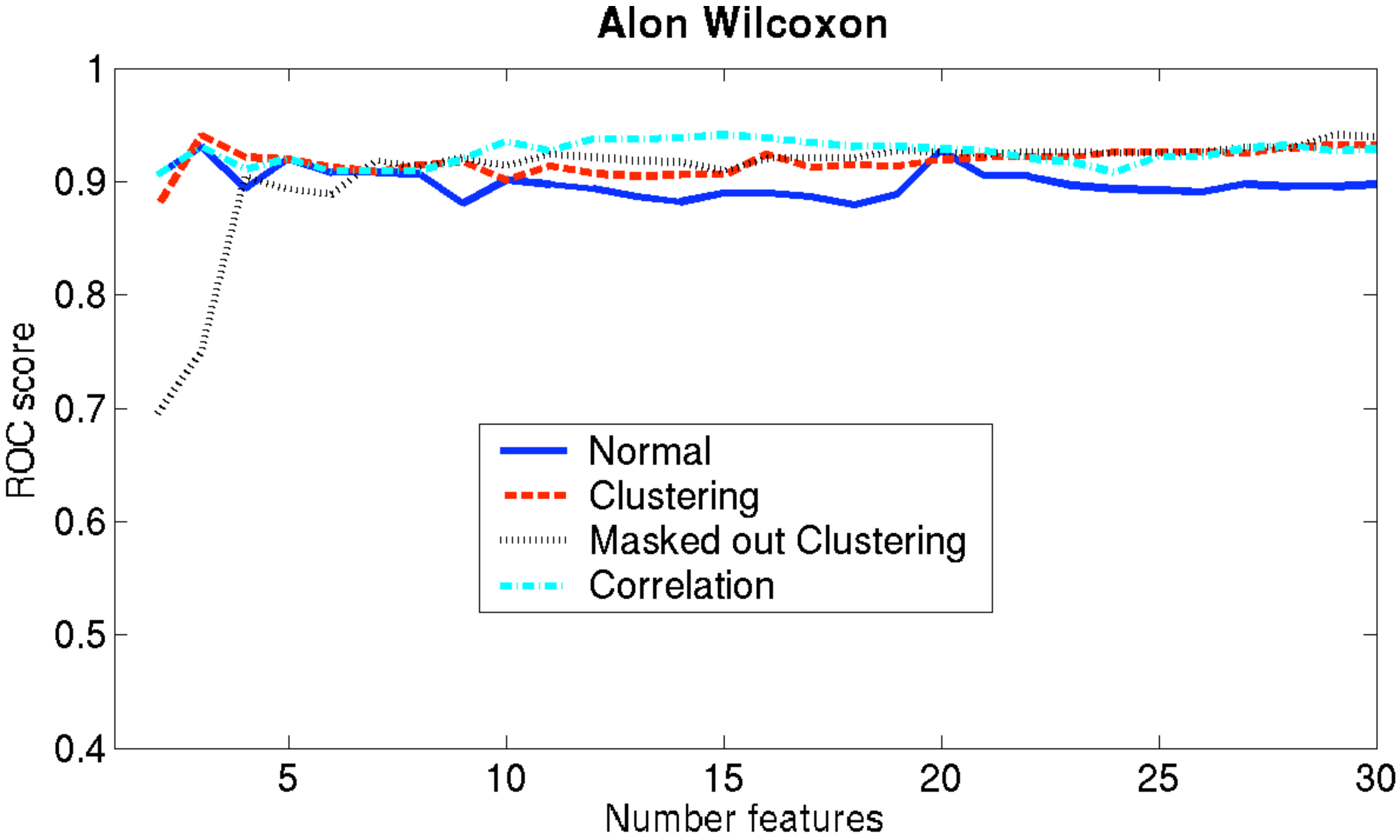
More ROC Scores



More ROC Scores



More ROC Scores



Overview

- Gene Expression Microarrays
- Classification and Feature Selection
- One Problem & Three Approaches
- Results
- → Summary and Conclusions

Summary I: Problem

- Sample classification is an important application of microarrays
  - For better diagnostics, prognostics, etc.
- Finding small feature sets with high classification accuracy is important
  - For cost, for biological insight
- "Standard" method (top k genes by your favorite statistical test) is not bad
  - But very often picks highly correlated subset

Summary II: Our Idea

- Explicitly pick subsets to emphasize diversity (reduced correlation) while retaining good individual statistics, hopefully will improve joint accuracy
- Three methods:
  - Greedy selection
  - Selection from clusters
  - Selection from clusters with masking

Summary III: Results

- It works
- Details vary a bit depending on data set and test statistic, but all 3 methods generally better than “standard”
- Improvement most significant for small feature set sizes
- Improvement greater for parametric tests than non-parametric tests

More Information

- Appeared in Pacific Symposium on Biocomputing, 2003
- Preprint, supplementary data
  - http://www.cs.washington.edu/homes/ruzzo
  - http://www.molgen.mpg.de/~jaeger/psb

Acknowledgements

- My coauthors
- Bill Noble
- Ranier Spang

- NIH
- NSF