



MEDIZINISCHE UNIVERSITÄT
INNSBRUCK

Microarray Analysis with Spherical Kernel Estimators using Correlation Information for Class Prediction

Florian Pedross

Department for Medical Statistics, Informatics
and Health Economics
Medical University of Innsbruck

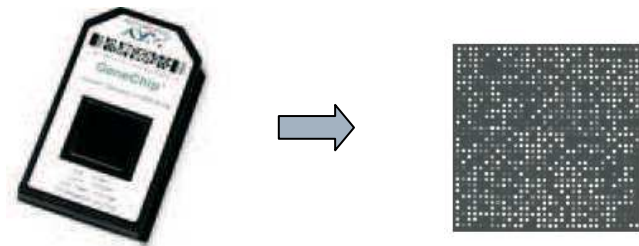
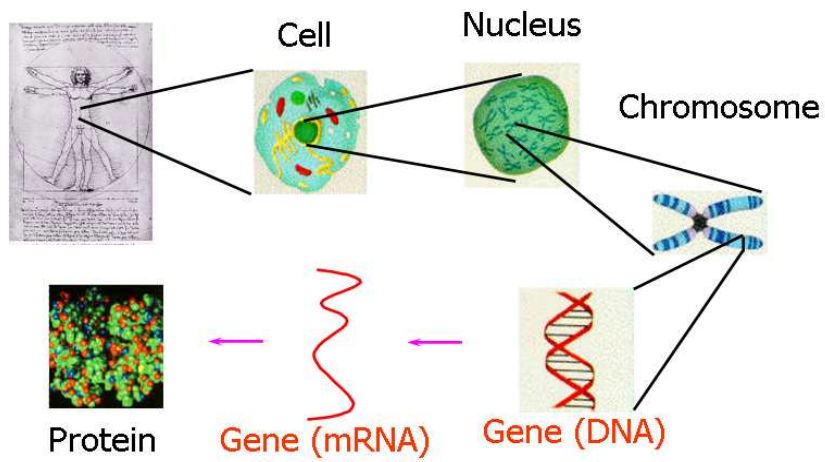
florian.pedross@i-med.ac.at

August 24, 2009

Overview

- Biological Background
- Class Prediction
- Kernel Estimator
- Gene Identification
- PDA
- Methods
- Results
- Conclusion

Biological Background



Statistical Methods for Microarrays

Different statistical questions arise:

1. Class discovery (unsupervised learning): clustering
2. Gene identification: selecting over-expressed or under-expressed genes
3. Class prediction (supervised learning):

Class Prediction

- Classifying observations into disjoint groups
- 2 Types:
 - Parametric Discriminant Analysis
 - Linear DA ($\ln(p(x | \omega_i)\pi_i)$)
 - Quadratic DA
 - Nonparametric Discriminant Analysis
 - Histogram
 - K-nearest neighbor
 - Naive estimator
 - Neural networks
 - Kernel estimator

Multidimensional Spherical Kernel

The random samples $\omega_{x_1}, \dots, \omega_{x_d}$ are given by matrix

$$X = \begin{pmatrix} x_{11} & \Lambda & x_{1d} \\ \mathbf{M} & \mathbf{O} & \mathbf{M} \\ x_{n1} & \mathbf{K} & x_{nd} \end{pmatrix}$$

Kernel Estimator:

$$\hat{f}(\bar{x}) = \frac{1}{n |H|} \sum_{i=1}^n K \left(H^{-1} \sqrt{(x - x_i)^T (x - x_i)} \right)$$

Multidimensional Gaussian Kernel

Properties:

- Continuous symmetric
- Kernel K not deciding
- Bandwidth (-matrix) important

Gaussian Kernel:

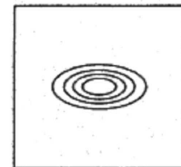
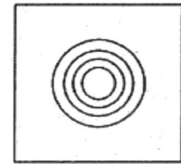
$$\hat{f}(\bar{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{(2\pi)^{d/2} |H|} \cdot \exp\left(-(\bar{x} - \bar{x}_i)^T H^{-1} (\bar{x} - \bar{x}_i)\right)$$

Bandwidth H

$$h_{opt} = \left(\frac{1}{n} \frac{\int K(z)^2 dz}{\int f''(z)^2 dz \cdot \int (z^2 K(z) dz)^2} \right)^{1/5}$$

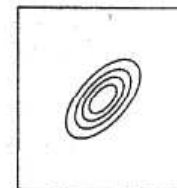
$$H = h \cdot I$$

$$H = \text{diag}(h_1, K, h_d)$$



Bandwidth proportional to covariance matrix:

$$H \propto \hat{\Sigma}^{1/2}$$



Gene Selection

- Using $H \propto \sum \hat{\sigma}^{1/2}$ no expensive method needed
- Regardless of correlation between genes
- Chi-square / Gain-Ratio
- Selected the (10) top ranked

Predictive Discrimination and Kernels

Bayes' Theorem $P(\omega_i | x) = \frac{p(x | \omega_i) \pi_i}{\sum_{i=1}^k \pi_i p(x | \omega_i)}$

where $p(x | \omega_i) = \frac{1}{n |H|} \sum_{i=1}^n K(H^{-1}(x - x_i))$

Classification: $p(x | \omega_g) \pi_g \geq p(x | \omega_k) \pi_k$

$\Rightarrow x \in \omega_g$

Data

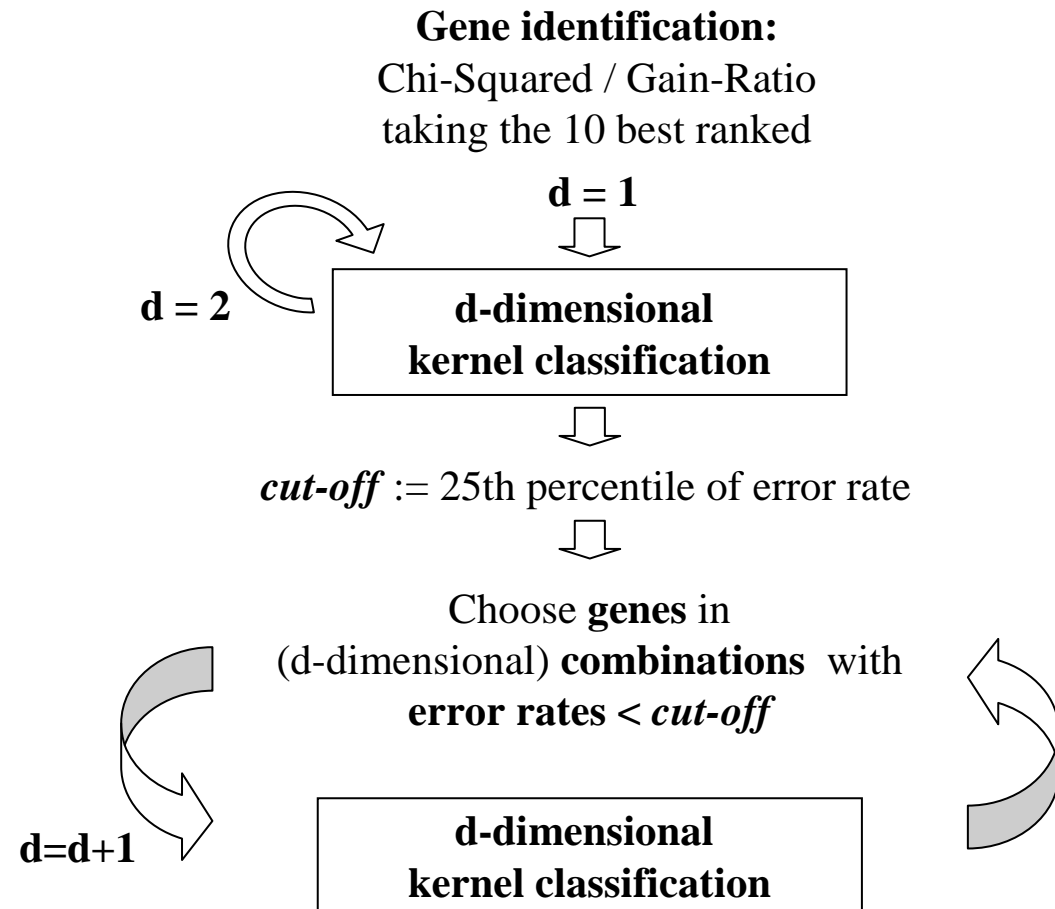
- Adrenocortical Tumors Data Set
 - 65 patients
 - 10,163 genes
 - 3 groups
 - 33 Adrenocortical Carcinoma
 - 22 Adrenocortical Adenoma
 - 10 Normal Adrenal Cortex
 - Data Source: GEO Accession: GSE10927

 - Breast Cancer Data Set
 - 62 patients
 - 10,163 genes
 - 2 groups
 - 43 Breast Tumor Profiles
 - 19 Normal Breast Tissue
 - Data Source: GEO Accession: GSE7904
-

□ Prostate Data Set

- 102 patients
- 4,487 genes
- 2 groups
 - 52 Prostate Tumor Profiles
 - 50 Normal Prostate Tissue
- Data Source: Broad Institute:
<http://www-genome.wi.mit.edu/MPR/prostate>

Methods



Results: Adrenocortical Tumors Data Set

Dimension	Probe-sets	R^2			
4		228268_at	227865_at	205911_at	222848_at
	228268_at	1			
	227865_at	0.61	1		
	205911_at	0.66	0.34	1	
	222848_at	0.67	0.63	0.47	1

Error rates:

$\propto \Sigma$	C - V	P - I
0.0	0.0308	0.0154

Results: Breast Cancer Data Set

Dimension	Probe-sets		R^2
2		202954_at	238062
	202954_at	1	
	238062_at	0.66	1

Error rates:

$\propto \Sigma$	C - V	P - I
0.0	0.0323	0.0323

Results: Prostate Data Set

Dimension	Probe-sets	R^2				
5		41706_at	37639_at	37720_at	38291_at	38634_at
	41706_at	1				
	37639_at	0.53	1			
	37720_at	0.22	0.44	1		
	38291_at	0.20	0.18	0.01	1	
	38634_at	0.24	0.19	0.01	0.42	1

Error rates:

$\propto \Sigma$	C - V	P - I
0.0388	0.1067	0.0679

Comparison with other Classifiers

Data set	Error Rates:			
	KE ($\propto \Sigma$)	SVM	NB	C4.5
ACC	0.0	0.0462	0.0615	0.1231
Breast	0.0	0.0161	0.0	0.0323
Prostate	0.0388	0.0784	0.0588	0.0980

Conclusion

- ❑ works with unbalanced data (contrast to Linear DA)
- ❑ no prior knowledge of a specific distribution needed
- ❑ small number of objects, large number of variables
- ❑ powerful results, especially by using the covariance bandwidth
- ❑ simple variable selection
- ❑ correlation can be used
- ❑ good discriminating genes can be lost
- ❑ long computing time

Literature - References

- Giordano T.J., Kuick R., Else T., Gauger P.G., Vinco M., Bauersfeld J., Sanders D., Thomas D.G., Doherty G. and Hammer G.(2009). **Molecular Classification and Prognostication of Adrenocortical Tumors by Transcriptome Profiling**. Clin Cancer Res, 15(2).
- Hastie T., Tibshirani R. and Friedman J. (2009). **The Elements of Statistical Learning**, New York: Springer
- Mitchell T.M. (1997). **Machine Learning, McGraw-Hill Series in Computer Science**, McGraw-Hill.
- Pedross F.A. (2008). **Nonparametric Predictive Discriminant Analysis Using Kernel Density Estimators**, Innsbruck
- Satagopan J.M. and Panageas K.S. 2003. **Tutorial in Biostatistics**, Statistics in Medicine, 22, 481-499.
- Silverman B.W. 1986. **Density Estimation for Statistics and Data Analysis**, London: Chapman & Hall.
- Singh D., Febbo P., Ross K., Jackson D., Manola J., Ladd C., Tamayo P., Renshaw A., D'Amico A., Richie J., et al.. 2002. **Gene expression correlates of clinical prostate cancer behaviour**, Cancer cell, I, 203-209.
- Webb A.R. 2002. **Statistical Pattern Recognition: Second Edition**, Chichester: Wiley & Sons

Thank You!

Contact

Dr. Florian Pedross
Department for Medical Statistics, Informatics
and Health Economics
Medical University of Innsbruck
Schöpfstrasse 41
6020 Innsbruck – Austria

Tel.: 0043 (0)512 9003 70903
email: florian.pedross@i-med.ac.at



MEDIZINISCHE UNIVERSITÄT
INNSBRUCK