GRC on Computer Aided Drug Design
Tilton, NH
July 31 – August 5, 2005

Partial Least Squares Support Vector Machines in Drug Design: hERG IC50 Analysis

Mark J. Embrechts and Curt M. Breneman
Departments of Decision Sciences and Engineering Systems, Chemistry, and Mathematical Sciences. Rensselaer Polytechnic Institute, Troy, NY

QSAR

- Predict a particular bio-activity from descriptors (attributes)
- Prediction process involves two stages which can be concurrent or distinct:
  - i) A feature reduction stage
    - GA-based feature selection
    - Sensitivity analysis
    - Kernel sigma-tuning based feature selection
  - ii) A predictive modeling stage
    - Artificial Neural Networks (ANN)
    - Support Vector Machines (SVM / LS-SVM)
    - Partially Least Square (PLS / K-PLS)
- Once relevant features have been identified, the predictive modeling stage is often straightforward and traditional.
- Feature selection for QSAR problems is a hard problem:
  - Inherent nonlinearity
  - Relatively large number of features
  - Small set of molecules with known bio-activities

hERG Dataset

- Millenium proprietary data provided by Dominique Ryan.
- 400 training data 109 test data.
- RECON, MOE and PEST descriptors

Comparative Analysis

descriptor	method	r2	R2	q2	Q2	sigma	#feats
RECON	K-PLS	0.75	0.75	0.468	0.476	7	147
PEST	K-PLS	0.97	0.97	0.515	0.516	15	896
MOE	K-PLS	0.82	0.82	0.587	0.616	7	189
REC/MOE	K-PLS	0.77	0.77	0.53	0.534	13	336
RECON	PLS	0.5	0.5	0.555	0.565		
PEST	PLS	0.61	0.61	0.552	0.568		
MOE	PLS	0.6	0.6	0.441	0.441		
REC/MOE	PLS	0.54	0.54	0.562	0.572		
RECON	DK-PLS	0.52	0.52	0.531	0.533		
PEST	DK-PLS	0.65	0.65	0.58	0.58		
MOE	DK-PLS	0.61	0.6	0.583	0.59		
REC/MOE	DK-PLS	0.5	0.5	0.564	0.568		
RECON	SVM	0.88	0.88	0.486	0.51		
PEST	SVM	0.998	0.991	0.52	0.526		
MOE	SVM	0.94	0.93	0.606	0.619		
REC/MOE	SVM	0.94	0.93	0.531	0.544		
RECON	LS-SVM	0.88	0.88	0.463	0.467		
PEST	LS-SVM	0.99	0.98	0.508	0.514		
MOE	LS-SVM	0.93	0.92	0.573	0.576		
REC/MOE	LS-SVM	0.92	0.9	0.535	0.538		

Descriptive Features

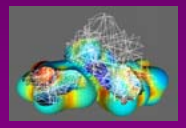
The Electronic Surface Properties

Recent investigations have revealed that a variety of intuitive chemical interaction mechanisms are well represented by specific patterns of surface property descriptors. These quantities may be approximated very rapidly using the RECON/TAE/PEST program, or may be computed more rigorously using the Gaussian98 or Jaguar programs.

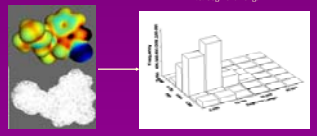
Orange regions represents one histogram bin

PEST-Shape Descriptors

- Surface property-encoding ray tracing
- TAE internal ray reflection – low resolution scan

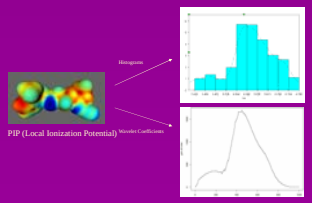


Isosurface (portion removed) with 750segments



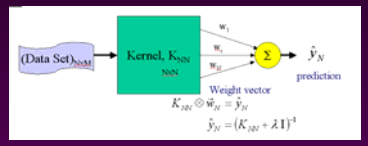
Shape-Aware Molecular Descriptors from Property/Segment-Length Distributions

Wavelet Representations of Surface Properties



Least Squares SVM

Ridge Regression with Kernel Trick



Equivalent SVM Formulation

$$\min_{w,b,e} J(w,b,e) = \sum_{i=1}^{N} e_i^2 + \lambda\, w^T w$$
$$s.t. \quad y_k\left[w^T \varphi(x_k) + b\right] = 1 - e_k, \quad k = 1,\ldots N$$

- Ridge Regression: Hoerl and Kennard (1970)
- Special case of Tikhonov regularization (1977)
- Formulated in dual/kernel context: Saunders Girosi/Poggio/Evangiou, Suykens, Mangasarian

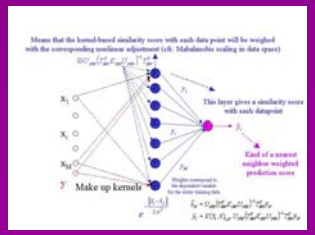
Ridge Regression, Regularization, LS-SVM

- λ ~ 1/C of normal SVMs
- Heuristic formula for λ (and also for SVM's C)
- Assumes data were first Mahalanobis scaled
- Verified on at least 40 different data sets

$$\lambda = \min\left[1, 0.05\left(\frac{N}{20}\right)^2\right] \quad (N \text{ is } \#\text{ of data points})$$

Data Set mws	PLS (lin) Q2_test	PLS (lin) LSME	P-SVM (lin) Q2_test	P-SVM (lin) LSME	K-PLS (rbf) Q2_test	K-PLS (rbf) LSME	K-PLS (cl2) Q2_test	K-PLS (cl2) LSME	DK-K-PLS Q2_test	DK-K-PLS LSME	LS-SVM Q2_test	LS-SVM LSME	P-SVM Q2_test	P-SVM LSME	SVMTorch Q2_test	SVMTorch LSME
	Time (s)		Time (s)		Time (s)		Time (s)		Time (s)		Time (s)		Time (s)		Time (s)	
Boston	0.28	0.28	**0.13**		0.14		0.18		0.14		0.18		0.16			
Housing	4.92	4.89	3.40		3.53		3.90		3.49		3.98		3.70			
506x13	1.35	1.18	181.05		158.31		**147.99**		167.60		661.49		212.43			
Albumin	0.36	0.42	0.35		0.41		**0.33**		0.35		**0.33**		0.38			
94x551	0.35	0.40	0.35		0.39		0.34		0.35		0.34		0.37			
	10.73	69.69	34.24		**10.36**		32.37		36.71		39.17		368.00			
Albalone	0.49	0.49	**0.44**		0.44		0.45		0.47		0.45		0.44			
4177x8	2.22	2.22	2.12		2.12		2.12		2.18		2.14		2.15			
	17.95	5.85	26224		24505		14202		**8298**		170925		6675			

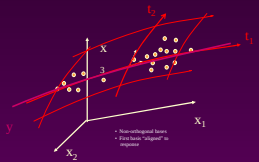
Least Squares PLS (K-PLS) in Data Space

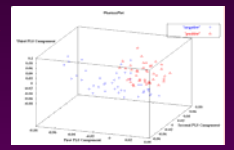


Partial Least Squares SVM

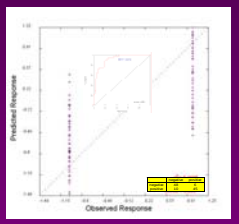
Peeking in SVM Space with K-PLS

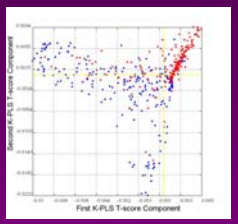
- Introduced by R. Rosipal and L. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space Space





Analysis with K-PLS





K-PLS Feature Selection with Sensitivity Analysis