Prediction by Nonparametric Posterior Estimation in Virtual Screening

Kitsuchart Pasupa (K.Pasupa@sheffield.ac.uk)
Department of Automatic Control & Systems Engineering
The University Of Sheffield.

Introduction

Virtual Screening (VS)

- Rank molecules so that most likely actives assayed first.

Binary Kernel Discrimination (BKD)

- Score $\propto$ Likelihood Ratio (LR) of active and inactive.
- Estimated by Parzen Windows.

Objectives

- Compute LR via direct estimate of posterior probability.
- Use non-parametric generalisation of logistic regression (Kernel Logistic Regression).
- Control complexity via Lq penalty function.

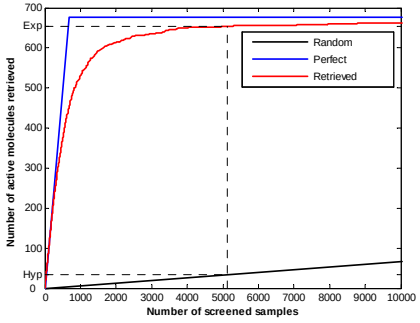

Fig. 1: Enrichment Plot

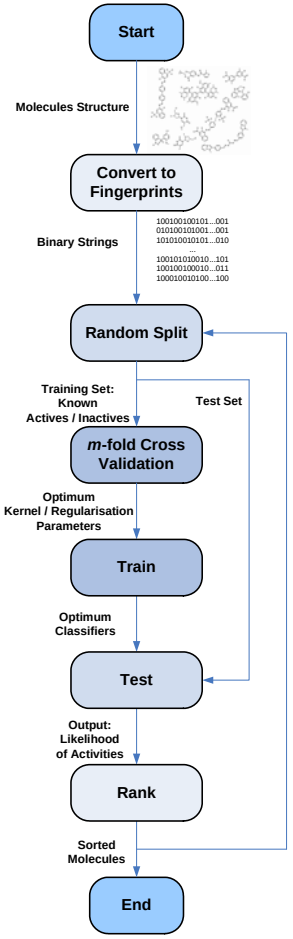

Fig. 2: Virtual Screening

References

[1] Chen, B., Harrison, R.F., Pasupa, K. Willett, P., Wilton, D.J., Wood, D.J., & Lewell, X.Q. J. (2006) Chem. Inf. Mod., 46:478-486.

[2] Zhu, J. & Hastie, T. (2005). J. Comp. Graph. Stat., 14, 185-205.

[3] Krishnapuram, B., Carin, L., Figueiredo, M.A. & Hartemink, A. (2005). IEEE Trans. Pattern Anal. Machine Intell., 27, 957-968.

Kernel Function

Binomial Kernel:

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \lambda^{n-d(\mathbf{x}_i,\mathbf{x}_j)}(1-\lambda)^{d(\mathbf{x}_i,\mathbf{x}_j)}$$

Hamming Distance (HD): Conventional

$$d(\mathbf{x}_i, \mathbf{x}_j) = b + c$$

Jaccard/Tanimoto Distance (J/T): [1]

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\frac{b+c}{a+b+c}\right) \cdot n$$

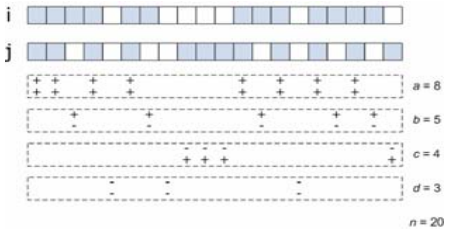

Fig. 3: Fingerprint

Kernel Logistic Regression

Logistic Model:

$$P(G_i \mid \mathbf{x}) = \frac{1}{1+\exp(-f(\mathbf{x},\boldsymbol{\omega}))}, i \in \{1,2\}$$

Linear Model: "Kernel trick"

$$f(\mathbf{x},\boldsymbol{\omega}) = \sum_{i=1}^{N} \omega_i k(\mathbf{x}, \mathbf{x}_i)$$

Minimise the Likelihood Function:

Kernel Logistic Regression (KLR)

$$\ell(\boldsymbol{\omega}) = \sum_{i=1}^{N} \log(1+\exp(-y_i f(\mathbf{x}_i))) + \rho N \|f(\mathbf{x})\|^2$$

Solved by using iteratively re-weighted least square (IRLS) algorithm.

Sparse Logistic Regression (SLR)

$$\ell(\boldsymbol{\omega}) = \sum_{i=1}^{N} \log(1+\exp(-y_i f(\mathbf{x}_i))) + \rho N \|f(\mathbf{x})\|_q^q$$

where, $0 < q \leq 1$

Solved by using EM-like algorithm

– Same computational complexity as IRLS.

– Solve non-convex problem.

KLR and BKD

$$LR \propto \frac{P(Active \mid \mathbf{x})}{P(Inctive \mid \mathbf{x})} = \frac{P(Active \mid \mathbf{x})}{1 - P(Active \mid \mathbf{x})} = \exp(f(\mathbf{x}, \boldsymbol{\omega}^*))$$

Experimental Results

>>MDL Drug Data Report (MDDR) database

>>1,024-D fingerprints representing 102,514 known drugs and biologically relevant molecules.

>>11 activity classes selected - reflect typical industrial drug discovery projects.

Idx	Activity Class	Self-Similarity Mean	Self-Similarity S.D.	BKD	KLR	SLR_1	$SLR_{0.5}$
1	Renin Inhibitors	0.337	0.105	99.10	**99.19**	98.51	98.29
	(226)					51.68	**11.95**
2	Angiotensin II AT1 Antagonists	0.269	0.100	97.43	**99.01**	98.47	98.37
	(190)					65.26	**24.84**
3	HIV Protease Inhibitors	0.226	0.101	**94.70**	93.69	92.80	93.60
	(150)					90.00	**77.33**
4	Thrombin Inhibitors	0.212	0.098	94.02	93.38	**94.04**	92.24
	(162)					76.79	**74.94**
5	Substance P Antagonists	0.179	0.082	**93.70**	93.52	90.99	91.90
	(250)					94.16	**85.28**
6	5HT3 Antagonists	0.175	0.090	**93.88**	93.29	90.64	90.87
	(150)					96.00	**82.27**
7	D2 Antagonists	0.173	0.089	**77.97**	76.85	73.30	68.73
	(80)					99.70	**71.50**
8	5HT1A Agonists	0.166	0.086	**88.28**	88.06	85.27	83.84
	(166)					96.75	**75.54**
9	5HTReuptake Inhibitors	0.153	0.092	**73.62**	73.19	69.16	67.00
	(72)					98.61	**85.28**
10	Protein Kinase C Inhibitors	0.141	0.103	**81.23**	79.66	75.68	71.74
	(92)					98.04	**74.57**
11	Cyclo-oxygenase Inhibitors	0.130	0.073	76.26	**76.78**	69.83	69.27
	(128)					99.69	**92.81**

Table 1: Self-Similarity, percent actives retrieved in top 5% of samples and its retained features

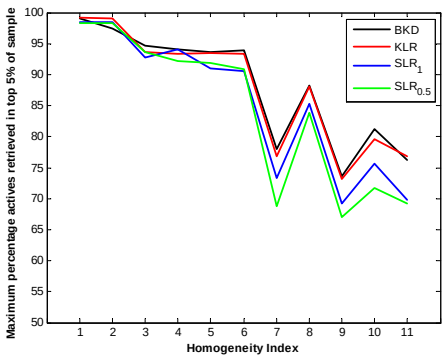

Fig. 4: Accuracy $\propto$ Homogeneity

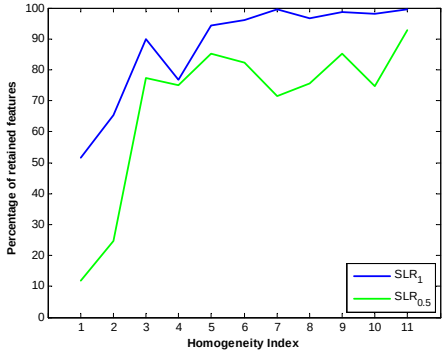

Fig. 5: Sparsity $\propto$ Homogeneity

Conclusions

>> Our BKD outperforms proposed method using J/T.

>> KLR & SLR only predict the probability of being Active.

>> Sparsity can be controllable but decreases in heterogeneous samples.

>> Sparseness desirable for high-through-put VS applications when computation of time is important.

2007 Symposium on Data Modelling for New Researchers
27th March 2007
University of Sheffield, UK