Interaction-Based Feature Selection and Classification for High-Dimensional Biological Data
Supplementary Data

Haitian Wang, Shaw-Hwa Lo, Tian Zheng, Inchi Hu

July 3, 2012

Supplement to Section 3.2.1

When choosing the size k of the initial subset of variables for BDA, the objective is to minimize the chance of erroneously dropping an influential variable, i.e. non-informative screening. If every element in the partition generated by the k variables of an initial subset contains at most one training case, the dropping is basically random. On the other hand, if there are partition elements with two or more training cases, the influence of a variable can then be informatively measured when dropped. Viewing elements of a partition as urns and training cases as balls, we evaluate the probability of at least 2 training cases in a partition element using an urn model as follows.

Let m be the number of urns and n be the number of balls. Dropping balls in urns randomly, let p_2 be the probability of two or more balls in a particular urn. The number of balls in a particular urn follows the binomial distribution with probability of success $1/m$ and the number of trials equals n. Thus

$$p_2 = 1 - \left(\frac{m-1}{m}\right)^n - \frac{n}{m}\left(\frac{m-1}{m}\right)^{n-1}.$$

Therefore mp_2 is the expected number of urns with 2 or more balls. We assume that $n \to \infty$ and $m \to \infty$ such that $n/m = \lambda$. Then

$$\begin{aligned} mp_2 &= m[1 - \left(\frac{m-1}{m}\right)^n - n\frac{1}{m}\left(\frac{m-1}{m}\right)^{n-1}] \\ &= m[1 - \left(\frac{m-1}{m}\right)^{m\frac{n}{m}} - \frac{n}{m}\left(\frac{m-1}{m}\right)^{m\frac{n-1}{m}}] \\ &\to m(1 - e^{-\lambda} - \lambda e^{-\lambda}) \end{aligned}$$

If we further assume that λ is small and close to zero, then

$$\begin{aligned} m(1 - e^{-\lambda} - \lambda e^{-\lambda}) &= m[1 - (1 - \lambda + \frac{\lambda^2}{2} + o(\lambda^2)) - \lambda(1 - \lambda + \frac{\lambda^2}{2} + o(\lambda^2))] \\ &= m[\frac{\lambda^2}{2} + o(\lambda^2)] \approx \frac{m}{2}\left(\frac{n}{m}\right)^2 = \frac{n^2}{2m} \end{aligned}$$

Therefore if we let m_{k-1} denote the number of elements in a partition generated by $k-1$ variables, then $n^2/2m_{k-1} \geq 1$, which is exact Equation (2) in the paper. Note that in reality the number of training cases in different partition elements are non-uniformly distributed and thus the expected number of elements with two or more training cases would be larger than 1. Therefore the initial size satisfying (2) represents a minimum requirement under uniform assumption.

A similar calculation for expected number of urns with three or more balls yields

$$\begin{aligned} & m(1-e^{-\lambda}-\lambda e^{-\lambda}-\frac{\lambda^2}{2}e^{-\lambda}) = me^{-\lambda}[e^{\lambda}-1-\lambda-\frac{\lambda^2}{2}))] \\ = \quad & m[\frac{\lambda^3}{3!}+o(\lambda^3)] \approx \frac{m}{6}\left(\frac{n}{m}\right)^3 = \frac{n^3}{6m_{k-1}^2} \end{aligned}$$

The corresponding condition on the starting size k is

$$n^3/6m_{k-1}^2 \geq 1. \tag{S.1}$$

When there are $n = 150$ training cases and each explanatory variable is binary, the largest integer k that satisfies (S.1) is 10 as opposed to 14 calculated from (2).

Supplement to Section 3.2.2

Consider the problem of placing a number of balls randomly into urns as described in Equation (3) of the paper. The following argument is similar to that on p. 113 of Aldous (1989). Assume the placement time follows a Poisson process of rate $\binom{k}{5}$. Then $P(\text{urn } j \text{ empty at time } t) \approx \exp\left[-t\binom{k}{5}/\binom{p}{5}\right]$ for a particular urn j. Poissonization makes urns independent. So $Q_t =$ the number of empty urns at time t satisfies $Q_t \approx$ Poisson with mean $\binom{p}{5}\exp\left[-t\binom{k}{5}/\binom{p}{5}\right]$. Let B be the first time all urns are occupied, then

$$P(B \leq t) = P(Q_t = 0) \approx \exp\left[-\binom{p}{5}\exp\left(-t\binom{k}{5}/\binom{p}{5}\right)\right].$$

This can be arranged to $\left[\binom{k}{5}/\binom{p}{5}\right]\left\{B - \left[\binom{p}{5}/\binom{k}{5}\right]\log\binom{p}{5}\right\} \approx \xi$; where $P(\xi \leq x) = \exp(-e^{-x})$. Therefore, $B \approx \left[\binom{p}{5}/\binom{k}{5}\right]\log\binom{p}{5}$ and Equation (3) in the paper is established. The generalization to clusters of size z is obtained by substituting 5 with z

$$B_z \approx \left[\binom{p}{z}/\binom{k}{z}\right]\log\binom{p}{z}.$$

In the toy example, we have $p = 30$ variables and would like to cover all quintuplets, then $B \approx \left[\binom{30}{5}/\binom{8}{5}\right]\log\binom{30}{5} \approx 2.5 \times 10^3$. If we repeat BDA $2B \approx 5000$ times, then we can expect to have a rather complete coverage of the quintuplets. Note that this is exactly the number of repetitions we used in the toy example.

Supplement to Section 3.2.3

The second filtering procedure adds one variable at a time to a return set and keep only those subsets with I-scores higher than that before adding, which are referred to as *forward-one sets*. Suppose there are H return sets after filtering out overlap ones, $\{R_h : h = 1, \ldots, H\}$. Let $|R_h|$ be the number of variables in R_h. Without loss of generality, assume $R_h = \{X_1, \ldots, X_{|R_h|}\}$. The remaining variables, $X_{|R_h|+1}, \ldots, X_p$ are added to R_h one at a time to generate $p - |R_h|$ subsets each of size $|R_h| + 1$. Let A_h be the number of forward-one sets, that is, the number of size $|R_h| + 1$ subsets by forward adding with I-scores higher than $I(R_h)$. If A_h is large, then R_h is removed.

The forward one procedure performs a kind of stability test on return sets. If a return set has no forward-one set ($A_h = 0$), then it is always returned by BDA whenever R_h is contained in the initial subset. The more forward-one sets a return set has, the more the return set depends on the

initial subset. For example, if $A_h = 7$, then there are 7 size $|R_h|+1$ initial subsets that will *not* lead to R_h if subjected to BDA. We can also replace the return set by a forward-one set if the later is of higher quality. Thus it also help in recovering influential variables missed in the previous stage.

From the data of van't Veer et al. (2002) , we selected two return sets from one of the 10 CV experiments. In Table S1, the 'false positive' return set has 7 forward-one sets even though its I-score is higher than the 'true' one, which has only 1 forward-one set. After removing the false positives, the final classification rule has considerably lower error rate in the test sample. Usually, the frequency plot of A_h has easily identified outliers and it is easy to determine a threshold on A_h to remove false positives. For example, in Figure S1, those return sets with 7 or more forward-one sets are considered false positives and should be removed.

Table S1: Examples of forward-one sets

	False		True	
Original return set	$\{665, 2283, 2930\}$	$I = 422.39$	$\{108, 2400, 4208\}$	$I = 410.778$
Forward-one sets subjected to BDA	$\{1451, 2930\}$	$I = 523.035$	$\{2930, 4208\}$	$I = 427.351$
	$\{665, 2283, 2930\}$	$I = 470.498$		
	$\{665, 1668, 2930\}$	$I = 450.050$		
	$\{1946, 2930\}$	$I = 438.888$		
	$\{1885, 2283, 2930\}$	$I = 438, 298$		
	$\{2283, 3291\}$	$I = 426.516$		
	$\{2283, 2900, 2930\}$	$I = 423.930$		

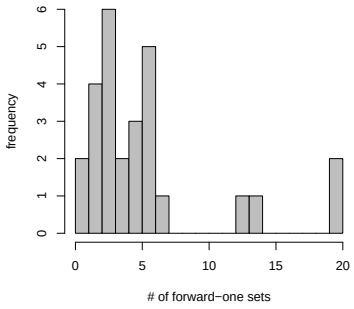

Figure S1: Frequency distribution for the number of forward-one sets

Supplement to Section 3.3.1

A sample output of logistic regression from R is shown in Exhibit 1 for a variable module of size 4 from the 78 training cases of vant Veer's data. After stepwise AIC selection, the final model dropped the 4-way and a 3-way interaction terms.

Exhibit 1. Model selection output using R on a module of 4 variables

```
Call:
glm(formula = y ~ g637 + g844 + g2145 + g3035 + g637:g844 + g637:g2145 +
g844:g2145 + g637:g3035 + g844:g3035 + g2145:g3035 + g637:g844:g3035 +
g637:g844:g2145 +g637:g2145:g3035, family = binomial(link =logit),
data = train)
Coefficients:
                   Estimate   Std.  Error  z value  Pr(>|z|)
 (Intercept)       5.8023     3.1413       1.847    0.0647 .
 g637              -0.2754    16.5963      -0.017   0.9868
 g844              4.6086     3.4669       1.329    0.1837
 g2145             -141.4701  61.3943      -2.304   0.0212 *
 g3035             -35.9947   20.6035      -1.747   0.0806 .
 g637:g844         -10.7009   25.8312      -0.414   0.6787
 g637:g2145        -533.0709  250.9616     -2.124   0.0337 *
 g844:g2145        -174.0768  76.3894      -2.279   0.0227 *
 g637:g3035        -169.9741  102.1318     -1.664   0.0961 .
 g844:g3035        -60.4414   31.1377      -1.941   0.0522 .
 g2145:g3035       80.1168    58.9844      1.358    0.1744
 g637:g844:g2145   -790.3717  353.9615     -2.233   0.0256 *
 g637:g844:g3035   -234.9399  155.3265     -1.513   0.1304
 g637:g2145:g3035  727.5103   356.0419     2.043    0.0410 *

Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)

Null deviance:  106.845 on 77 degrees of freedom
Residual deviance:  41.126 on 64 degrees of freedom
AIC: 69.126
Number of Fisher Scoring iterations:  11
```

Supplement to Section 3.3.2

Exhibition 2. Boosting Algorithm for Variable Modules

1. Input H variable modules.
2. Initialize training-case weights $w_i = 1/n,\ i = 1, \cdots, n$.
3. For $h = 1$ to H do
   (a) For $j = h$ to H, fit logistic regression classifier L_j to the training data using variable module R_k. Calculate
   $$\text{err}_j = \frac{\sum_i w_i I(y_i \neq L_j(x_i))}{\sum_i w_i}, \quad \text{and } \alpha_j = \frac{1}{2}\log\frac{1-\text{err}_j}{\text{err}_j}. \tag{S.2}$$
   (b) Let $j' = \underset{h \leq j \leq H}{\text{argmax}}\ \alpha_j$. Update $\quad w_i \leftarrow w_i \times \exp(\alpha_{j'} I(y_i \neq L_{j'}(x_i)))$.
   (c) Relabel $R_{j'}$ as R_h with corresponding α_h. And relabel the remaining $H - h$ variable modules as $\{R_{h+1}, \ldots, R_H\}$.
4. Output the classification rule: $\text{sign}\{\sum_{h=1}^{H} \alpha_h L_h\}$.

Supplement to Section 4.1

Table S2: Error rates on van't Veer dataset by various methods

Author	Feature selection	Classifier evaluated	Error rate	Method
Pochet et al (2004)	None	LS-SVM[a]	0.310	LOOCV
		Linear kernel	0.321	Test set
	None	LS-SVM	0.309	LOOCV
		RBF[b] kernel	0.316	Test set
	None	LS-SVM	0.478	LOOCV
		no regularization	0.428	Test set
	PCA[c]	LDA	0.297	LOOCV
	(unsupervised)		0.426	Test set
	PCA	LDA	0.265	LOOCV
	(supervised)		0.331	Test set
	PCA linear kernel	LDA	0.288	LOOCV
	(unsupervised)		0.391	Test set
	PCA linear Kernel	LDA	0.264	LOOCV
	(supervised)		0.346	Test set
	PCA RBF kernel	LDA	0.251	LOOCV
	(unsupervised)		0.486	Test set
	PCA RBF kernel	LDA	0.000	LOOCV
	(supervised)		0.632	Test set
Li and Yang (2005)	RFE	SVM	0.105⋆	Test set
	RFE	Ridge regression	0.158⋆	Test set
	RFE	Rocchio	0.158⋆	Test set
Michiels et al (2005)	Correlation	Correlation	0.310	500 rCV[k]
Peng (2005)	Golub[d]	SVM	0.247⋆	LOOCV
	Golub	Bagging SVM	0.226⋆	LOOCV
	Golub	Boosting SVM	0.226⋆	LOOCV
	Golub	Ensemble SVM	0.186⋆	LOOCV
Yeung et al (2005)	BMA[e]	BMA	0.158⋆	Test set
Alexe et al (2006)	LAD[f]	LAD	0.183⋆	CV
Diaz-Uriarte and	None	Random forest	0.342	bootstrap
de Anres (2006)	None	SVM	0.325	bootstrap
	None	kNN[g]	0.337	bootstrap
	None	LDA	0.331	bootstrap
	Shrunken centroid	Shrunken centroid	0.324	bootstrap
	NNVS[h]	NNVS	0.337	bootstrap
Wahde & Szallasi (2006)	Evolutionary algorithm	LDA	0.105⋆	Test set
Song et al (2007)	RFE	SVM	0.077⋆	10-fold CV
Yan and Zheng (2008)	sMPAS[i]	sMPAS	0.295	13-fold CV
Zhu et al (2008)	RFE	SVM	0.29	10-fold CV
Liu et al (2009)	EICS[j]	EICS	0.219	10-fold rCV
The proposed method	Retention frequency	Boosting logistic	**0.080**	10-fold rCV
			0.000	Test set

[a]Least square SVM; [b]Radial basis function; [c]Principle component analysis; [d]The feature selection method in Golub et al (1999); [e]Bayesian model averaging; [f]Logical analysis of data; [g]k-nearest neighbor; [h]Nearest neighbor with variable selection; [i]Signed multigene association; [j]Ensemble independent component system; [k]Random CV; ⋆Biased estimates due to turning parameter selection and/or feature selection

Supplement to Section 4.2

Table S3: Biological implication of identified genes in van't Veer data

Gene module		Systematic name	Gene name	Description
I	1	Contig45347_RC	KIAA1683	ESTs
	2	NM_005145	GNG7	guanine nucleotide binding protein (G protein), gamma 7
	3	Z34893	**ICAP-1A**	integrin cytoplasmic domain-associated protein 1
	4	NM_006121	KRT1	keratin 1 (epidermolytic hyperkeratosis)
	5	NM_004701	**CCNB2**	cyclin B2
II	1	AB007950	KIAA0481	KIAA0481 gene product
	2	Contig53226_RC		ESTs
	3	Contig12369_RC		ESTs
	4	NM_006806	**BTG3**	BTG family, member 3
III	1	NM_003303	TRO	trophinin
	2	NM_002809	PSMD3	proteasome (prosome, macropain) 26S subunit, non-ATPase, 3
	3	NM_014176	HSPC150	HSPC150 protein similar to ubiquitin-conjugating enzyme
	4	NM_016077	LOC51651	CGI-147 protein
IV	1	NM_013232	PDCD6	programmed cell death 6
	2	NM_005375	**MYB**	v-myb avian myeloblastosis viral oncogene homolog
	3	NM_018182	FLJ10700	hypothetical protein FLJ10700
	4	Contig39950_RC		ESTs
V	1	NM_004119	FLT3	fms-related tyrosine kinase 3
	2	NM_020675	AD024	Homo sapiens AD024 protein (AD024), mRNA.
	3	NM_016632	LOC51326	ARF protein
	4	Contig53912_RC		Homo sapiens mRNA; cDNA DKFZp547M146 (from clone DKFZp547M146)
	5	Contig49670_RC		Homo sapiens cDNA: FLJ23228 fis, clone CAE06654
VI	1	NM_003087	**SNCG**	synuclein, gamma(breast cancer-specific protein 1)
	2	D38553	KIAA0074	KIAA0074 protein
	3	NM_001216	**CA9**	carbonic anhydrase IX
VII	1	NM_001741	CALCA	calcitonin/calcitonin-related polypeptide, alpha
	2	NM_005132	REC8	Rec8p, a meiotic recombination and sister chromatid cohesion phosphoprotein of rad21p family
	3	Contig46_RC		ESTs
	4	NM_000427	**LOR**	loricrin

Gene module		Systematic name	Gene name	Description
VIII	1	NM_019854	HRMT1L3	HMT1 (hnRNP methyltransferase, S. cerevisiae)-like 3
	2	Contig20816_RC		ESTs
	3	Contig55377_RC		ESTs
IX	1	Contig45816_RC		ESTs
	2	NM_020411	**XAGE-1**	XAGE-1 protein
	3	AB004064	TMEFF2	transmembrane protein with EGF-like and two follistatin-like domains 2
	4	Contig34634_RC	GCN1L1	GCN1 (general control of amino-acid synthesis 1, yeast)-like 1
X	1	NM_020166	**MCCC1**	3-methylcrotonyl-CoA carboxylase biotin-containing subunit
	2	NM_012261	HS1119D91	similar to S68401 (cattle) glucose induced gene
	3	NM_018265	FLJ10901	hypothetical protein FLJ10901
	4	Contig39090_RC		ESTs
XI	1	Contig53968_RC		ESTs
	2	NM_004774	PPARBP	PPAR binding protein
	3	NM_007117	TRH	thyrotropin-releasing hormone
	4	NM_000599	**IGFBP5**	Homo sapiens insulin-like growth factor binding protein 5 (IGFBP5), mRNA.
XII	1	NM_016359	LOC51203	clone HQ0310 PRO0310p1
	2	Contig41383_RC		ESTs
XIII	1	NM_004603	**STX1A**	syntaxin 1A (brain)
	2	AB020713	KIAA0906	KIAA0906 protein
	3	NM_000231	SGCG	sarcoglycan, gamma (35kD dystrophin-associated glycoprotein)
XIV	1	Contig52018_RC		ESTs
	2	Contig19224_RC		ESTs
	3	NM_018304	FLJ11029	hypothetical protein FLJ11029
	4	NM_002196	INSM1	insulinoma-associated 1
XV	1	NM_004791	**ITGBL1**	integrin, beta-like 1 (with EGF-like repeat domains)
	2	AF055033	**IGFBP5**	insulin-like growth factor binding protein 5
	3	NM_006681	**NMU**	neuromedin U
XVI	1	Contig34964_RC		ESTs
	2	NM_012177	**FBXO5**	F-box only protein 5
	3	Contig55181_RC		ESTs
XVII	1	NM_004994	**MMP9**	matrix metalloproteinase 9 (gelatinase B, 92kD gelatinase, 92kD type IV collagenase)
	2	AK001100	**DSC3**	Homo sapiens cDNA FLJ10238 fis, clone HEMBB1000449
XVIII	1	NM_004163	**RAB27B**	RAB27B, member RAS oncogene family
	2	Contig55829_RC		ESTs
	3	Contig173		ESTs

Supplement to Section 4.3.2

Table S4: Biological implication of identified genes in Golub data

Gene module		Systematic name	Gene name	Description
I	1	X16323_at	HGF	HGF Hepatocyte growth factor (hepapoietin A; scatter factor)
	2	D86961_at	LHFPL2	lipoma HMGIC fusion partner-like 2
	3	Y12670_at	LEPROT	LEPR Leptin receptor
	4	D87074_at	RIMS3	regulating synaptic membrane exocytosis 3
	5	D26308_at	BLVRB	biliverdin reductase B (flavin reductase (NADPH))
II	1	U04898_at	RORA	RAR-related orphan receptor A
	2	M58297_at	MZF1	ZNF42 Zinc finger protein 42 (myeloid-specific retinoic acid-responsive)
	3	J03473_at	PARP1	ADPRT ADP-ribosyltransferase (NAD+; poly (ADP-ribose) polymerase)
III	1	D87078_at	PUM2	pumilio homolog 2 (Drosophila)
	2	D83785_at	MAML1	mastermind-like 1 (Drosophila)
IV	1	D86983_at	PXDN	VLDLR Very low density lipoprotein receptor
	2	U14603_at	PTP4A2	protein tyrosine phosphatase type IVA
	3	M81933_at	CDC25A	cell division cycle 25 homolog A (S. pombe)
	4	X77307_at	HTR2B	5-hydroxytryptamine (serotonin) receptor
	5	D16532_at	VLDLR	very low density lipoprotein receptor
V	1	U35451_at	CBX1	chromobox homolog, Heterochromatin protein p25 mRNA
	2	M31551_s_at	SERPINB2	serpin peptidase inhibitor, clade B (ovalbumin)
	3	M55150_at	FAH	fumarylacetoacetate
VI	1	HG1496-HT1496_s_at		Adrenal-Specific Protein Pg2
	2	U12471_cds1_at	THBS1	thrombospondin 1
	3	M23197_at	CD33	CD33 antigen (differentiation antigen)
	4	X03934_at	CD3D	CD3d molecule, delta (CD3-TCR complex)
	5	U59877_s_at	RAB31	member RAS oncogene family Rab22B
VII	1	M38690_at	CD9	CD9 molecule GIG2
	2	X00437_s_at	TRBC1	T cell receptor beta constant 1
VIII	1	M12759_at	IGJ	immunoglobulin J polypeptide, linker protein for immunoglobulin alpha and mu polypeptides
	2	M23323_s_at	CD3E	t-cell surface glycorpotein epsilon chain precursor
	3	X52142_at	CTPS	CTP synthetase
	4	X59417_at	KIAA039	proteasome iota chain
IX	1	U22376_cds2_s_at	MYB	v-myb myeloblastosis viral oncogene homolog (avian)
	2	U90902_at	TIAM1	T-cell lymphoma invasion and metastasis 1
	3	U91903_at	FRZB	frizzled-related protein FRE

Gene module		Systematic name	Gene name	Description
X	1	J04430_s_at	ACP5	acid phosphatase 5, tartrate resistant
	2	S68805_at	GATM	glycine amidinotransferase(L-arginine: glycine amidinotransferase)
	3	U14193_at	GTF2A2	general transcription factor IIA
XI	1	Y00339_s_at	CA2	carbonic anhydrase II
	2	X59350_at	CD22	CD22 antigen
	3	X59871_at	TCF7	transcription factor 7 (T-cell specific, HMG-box)
XII	1	M19888_at	SPRR1B	small proline-rich protein 1B
	2	U09413_at	ZNF135	Zinc finger protein 135 (clone pHZ-17)
	3	M28170_at	CD19	CD19 antigen
XIII	1	X69111_at	ID3	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein
	2	Z49148_s_at	RPL29P11	ribosomal protein L29 pseudogene 11
XIV	1	M84371_rna1_s_at	CD19	CD19 molecule B4
	2	M83652_s_at	CFP	complement factor properdin
XV	1	X02874_at	OAS1	(2'-5') oligoadenylate synthetase 1
	2	D00749_s_at	CD7	T-cell antigen CD7 precursor
	3	U23852_s_at	LCK	T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA
	4	U90552_at	BTN3A1	butyrophilin, subfamily 3, member A1
XVI	1	L08010_at	REG1B	regenerating islet-derived 1 beta
	2	X58288_at	PTPRM	protein tyrosine phosphatase, receptor type, mu polypeptide
	3	HG2479-HT2575_s_at		Helix-Loop-Helix Protein Sef2-1d

References

[1] Aldous, D. (1989), *Probability Approximation via Poisson Clumping Heuristic*, Springer-Verlag, New York.

[2] Alexe, B., Alexe, S., Axelrod, D. et al (2006), Breast cancer prognosis by combinatorial analysis of gene expression data, *Breast Cancer Research*, **8:R41**.

[3] Diaz-Uriate, R. and de Andres, S. A. (2006), Gene selection and classification of microarray data using random forest, *BMC Bioinformatics*, **7:**3.

[4] Li, F. and Yang, Y. M. (2005), Analysis of recursive gene selection approaches from microarray data, *Bioinformatics*, **21**(19), 3741-3747.

[5] Liu, K. H. *et al* (2009), Microarray data classification based on ensemble independent component selection, *Computers in Biology and Medicine*, **39**, 953-960.

[6] Michiels, S. S. *et al.* (2005), Prediction of cancer outcome with microarrays: a multiple random validation strategy, *Lancet*, **365**(9458), 488-492.

[7] Peng, Y. H. (2005), Robust ensemble learning for cancer diagnosis based on microarray classification, *Adv. Data Min. and Appl., Proc.*, **3584**, 564-574.

[8] Pochet, N. F. *et al* (2004), Systematic benchmarking of microarray and classification: assessing the role of non-linearity and dimensionality reduction, *Bioinformatics*, **20**(17), 3185-3195.

[9] Song, L., Bedo, J., Borgwardt, K.M., et al (2007), Gene selection via the BAHSIC family of algorithms, *Bioinformatics (ISMB)*, **23**(13), i490-i498.

[10] van't Veer, L. J. *et al* (2002), Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530 - 536.

[11] Wahde, M. and Szallasi, Z. (2006), Improving the prediction of the clinical outcome of breast cancer using evolutionary algorithms, *Soft Comp.*, **10**(4), 338-345.

[12] Yan, X. and Zheng, T. (2008), Selecting informative genes for discriminant analysis using multigene expression profiles, *BMC Gen.*, **9**(Spl. 2), 1471-2164-9-S2-S14.

[13] Yeung, K. Y., *et al.* (2005), Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data, *Bioinformatics*, **21**(10), 2394-2402.

[14] Zhu, J. X. *et al.* (2008), On selection bias with prediction rules formed from gene expression data , *J. Stat. Plann. and Infer.*, **138**, 374-386.