Mixture of experts to combine clinical factors and gene markers

Kim-Anh Lê Cao

ARC Centre of Excellence in Bioinformatics
&
Queensland Facility for Advanced Bioinformatics
The University of Queensland

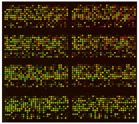
QFAB KNOWLEDGE FROM DATA

Australian Government
Australian Research Council

The University Of Queensland
AUSTRALIA

Microarray vs. clinical data

Microarray data



- generate insight into cell biology
- identify marker genes to predict prognosis
- complex and noisy nature
- few validated biomarkers

Clinical factors

- valuable information
- low noise level
- used as prognosis factors but considered not sufficient to predict patient outcome

Aim

Clinical data and gene expression data both contain complementary information for cancer prognosis and therapeutic targeting.

Integrating both types of data:

→ may lead to a more powerful prognosis prediction (improvement in the accuracy)

→ may help reduce the number of marker genes to reliably predict the prognosis.

Statistical challenges

Clinical variables often are

- categorical
- heterogeneous (ER +/- status, histological grade, age, ...)

Gene expression variables are

- continuous variables
- homogeneous

→ not easily combined in a classification approach !

Related litterature

Few statistical methodologies proposed and little success so far ...

e.g. on Van' t Veer breast cancer data set:

- 👎 Edén et al. (ANN, 2004), Dettling and Buhlmann* (2004, PELORA), Boulesteix et al.* (2008, PLS-RF)
- 👉 Gevaert et al. (2006, Bayesian networks)
- 👍 Sun et al. (2007, I-RELIEF)

→ depends on the statistical approach

→ depends on the data set

→ few approaches deal with *categorical* clinical factors (*)

Integrative Mixture of Experts

1. Select the relevant genes
2. Combine both types of variables using mixture of experts
3. Assess the biological relevance of the selected genes

→ Application to three cancer data sets

Gene selection

Genes are selected based on the outcome status using 10 fold cross-validation with three types of gene selection procedures:

- univariate filter approach: t-test
- wrapper approach: Random Forests (Breiman, 2001)
- sparse PLS-DA (sPLS, Lê Cao et al., 2008, 2009a, `integrOmics`, 2009b)

Mixture of Experts

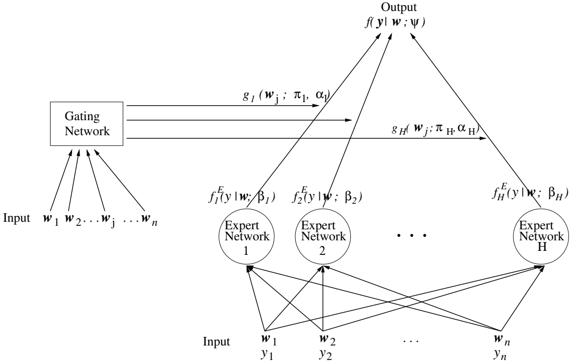
Mixture of experts models (ME, Jacobs et al., 1991)

- account for nonlinearities and other complexities in the data
- based on a divide-and-conquer strategy
- wide applicability
- advantages of fast learning via EM algorithm

Mixture of Experts were improved

- for classification problems (Ng & McLachlan, 2007)
- *integrative ME*: deals with categorical and continuous variables together (Ng & McLachlan, 2008)

Mixture of Experts


Output
$f(\mathbf{y}|\mathbf{w};\psi)$
Gating Network
$g_1(\mathbf{w}_j;\pi_1,\alpha_1)$
$g_H(\mathbf{w}_j;\pi_H,\alpha_H)$
Input
$\mathbf{w}_1\ \mathbf{w}_2 \dots \mathbf{w}_j\ \dots \mathbf{w}_n$
$f_1^E(y|\mathbf{w};\beta_1)$
$f_2^E(y|\mathbf{w};\beta_2)$
$f_H^E(y|\mathbf{w};\beta_H)$
Expert Network 1
Expert Network 2
. . .
Expert Network H
Input
$\mathbf{w}_1$ y_1
$\mathbf{w}_2$ y_2
. . .
$\mathbf{w}_n$ y_n

$\mathbf{y}_j$: outcome of patient j
$\mathbf{x}_j$: gene signature
$\mathbf{z}_j$: clinical factors
$\mathbf{w}_j = (\mathbf{x}_j^T, \mathbf{z}_j^T)^T$: hybrid signature

Both experts and gating networks receive $\mathbf{w}_j$ as input.
Final output is a linear combination of the expert and gating networks' outputs.

Mixture of Experts

- Expert network: each input is modeled via a Bernoulli distribution

$$f_h^E(y_j|\boldsymbol{w}_j;\beta_h) = \left(\frac{\exp(\beta_h^T\boldsymbol{w}_j)}{1+\exp(\beta_h^T\boldsymbol{w}_j)}\right)^{y_j}\left(\frac{1}{\exp(\beta_h^T\boldsymbol{w}_j)}\right)^{(1-y_j)}$$

- Gating network: different types of gating functions are proposed

$$g_h(\boldsymbol{w}_j;\boldsymbol{\pi}_h,\boldsymbol{\alpha}_h) = \frac{\boldsymbol{\pi}_h f_h^G(\boldsymbol{w}_j;\boldsymbol{\alpha}_h)}{\sum_{l=1}^H \boldsymbol{\pi}_l f_l^G(\boldsymbol{w}_j;\boldsymbol{\alpha}_h)}$$

- Final output: weighted sum of all the local output vectors produced by the experts and the gating network

$$f(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\Psi}) = \sum_{h=1}^H g_h(\boldsymbol{w};\boldsymbol{\pi}_h,\boldsymbol{\alpha}_h)f_h^E(y|\boldsymbol{w};\beta_h)$$

Application of Mixture of Experts

Gating function

$$g_h(\boldsymbol{w}_j;\pi_h,\alpha_h)=\frac{\pi_h f_h^G(\boldsymbol{w}_j;\alpha_h)}{\sum_{l=1}^H \pi_l f_l^G(\boldsymbol{w}_j;\alpha_h)}$$

- Multinomial logit model
- Independent model (Ng & McLachlan, 2008)?
- Location model (Hunt & Jorgensen, 1999)

→ fitted with EM algorithm

Data sets

	p	q	No. of Samples class 0	 class 1	Ref.
Prostate	7,884	8	37 (rec)	42 (no rec)	Stephenson et al. (2005)
Breast	5,537	8	75 (rec)	181 (no rec)	van de Vivjer et al. (2002)
CNS	7,128	5	21 (dead)	39 (alive)	Pomeroy et al. (2002)

p: the number of transcripts, q: the number of clinical factors.

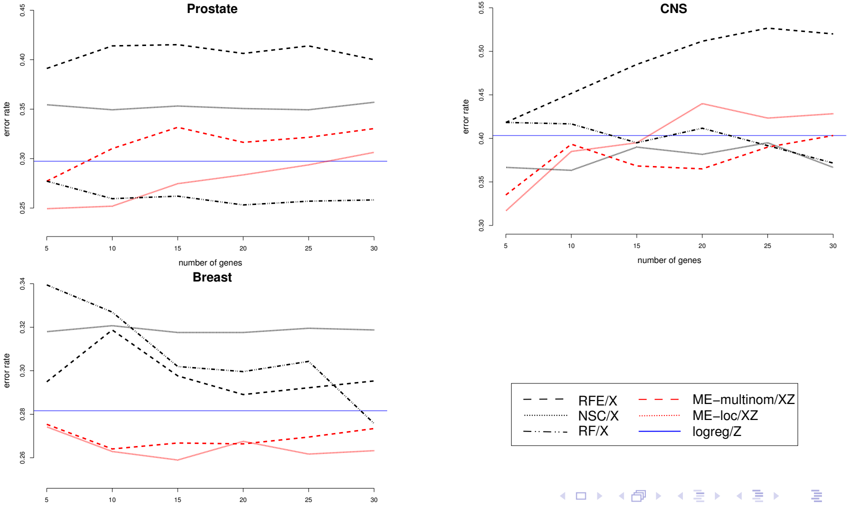
→ careful use of cross-validation during gene selection step

→ integrative ME is learnt on a training set and prediction is evaluated on a test set

Assessing additional predictive value

1. On the gene expression data alone
   Wrapper approaches perform internal variable selection:
   - Recursive Feature Elimination (RFE, Guyon et al. 2002)
   - Nearest Schrunken Centroids (NSC, Tibshirani et al. 2002)
   - Random Forests (RF, Breiman 2001)
2. On the clinical data alone
   - Logistic regression
3. On gene expression and clinical data
   Integrative ME with different gating functions:
   - Multinomial logit
   - Location model

Error rate estimation: ME + t-test



In a nutshell

- integrative ME is more accurate than clinical variables alone
- integrative ME is often more accurate than microarray data alone especially when the number of genes is small
- performance also depends on the data set

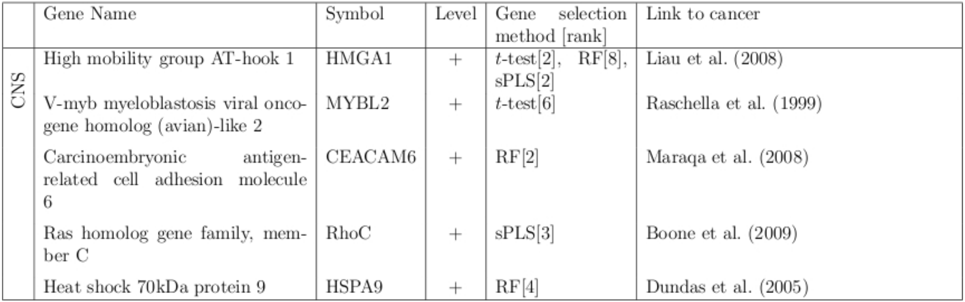
Link with biology ?

- Is the proposed hybrid signature biologically relevant ?
- Is there any difference between the gene selection procedures ?

Biological relevance: Prostate & Breast cancers

	Gene Name	Symbol	Level	Gene selection method [rank]	Link to cancer
Prostate	Etoposide induced 2.4 mRNA	EI24	+	*t*-test[1], RF[1], sPLS[1]	Gu et al. (2000); Zhao et al. (2005)
	Erythrocyte membrane protein band 4.9	EPB49	-	*t*-test[2], sPLS[1]	Lutchman et al. (1999)
	Chromatin modifying protein 1A	CHMP1A	-	*t*-test[5], RF[2], sPLS[5]	Li et al. (2008)
	Asparagine synthetase	ASNS	+	RF[4]	Richards and Kilberg (2006); Estes et al. (2007)
	Prothymosin, alpha	PTMA	+	RF[5]	Suzuki et al. (2006)
Breast	Insulin-like growth factor binding protein 5	IGFBP5	+	*t*-test[1,3], RF[5,8,13], sPLS[1,3]	Nishidate et al. (2004); van't Veer et al. (2002); Li et al. (2007); Mita et al. (2007)
	Phosphoglycerate mutase 1	PGK1	+	*t*-test[2], RF[11], sPLS[2]	Duan et al. (2002); Hwang et al. (2006); Zhang et al. (2005); Zieker et al. (2008)
	Protein regulator of cytokinesis 1	PRC1	+	*t*-test[5], RF[12], sPLS[5]	Shimo et al. (2007)
	E2F transcription factor 1	E2F1	+	*t*-test[13]	Zhang et al. (2000); Vuaroqueaux et al. (2007)
	Adrenomedullin	ADM	+	RF[6]	Oehler et al. (2003)

Biological relevance: CNS cancer



	Gene Name	Symbol	Level	Gene selection method [rank]	Link to cancer
CNS	High mobility group AT-hook 1	HMGA1	+	*t*-test[2], RF[8], sPLS[2]	Liau et al. (2008)
	V-myb myeloblastosis viral oncogene homolog (avian)-like 2	MYBL2	+	*t*-test[6]	Raschella et al. (1999)
	Carcinoembryonic antigen-related cell adhesion molecule 6	CEACAM6	+	RF[2]	Maraqa et al. (2008)
	Ras homolog gene family, member C	RhoC	+	sPLS[3]	Boone et al. (2009)
	Heat shock 70kDa protein 9	HSPA9	+	RF[4]	Dundas et al. (2005)

Different gene selection approaches often highlight different genes

→ relevant and complementary information

→ potential biomarkers need to be further validated

Conclusion

- Noisy characteristic of gene expression data can be compensated by clinical variables
- Both types of variables are useful to predict cancer prognosis
- Integrative ME is a sound approach and can deal with continuous and categorical variables
- Biologically relevant results were obtained
- R package `integrativeME`
- Improvements with larger-scale studies involving the records of a larger number of clinical variables

Acknowledgements

QFAB
KNOWLEDGE FROM DATA

Prof. Geoff. McLachlan	Univ. QLD
Dr. Emmanuelle Meugnier	Univ. Lyon
Dr. Shu-Kay Ng	Griffith University

The University
Of Queensland
AUSTRALIA

Australian Government
Australian Research Council

Merci pour votre attention !

k.lecao@uq.edu.au