Minimum Number of Genes for Microarray Feature Selection

Elena Baralis, Giulia Bruno, Alessandro Fiori

{elena.baralis, giulia.bruno, alessandro.fiori}@polito.it

Dipartimento di Automatica e Informatica

Politecnico di Torino

Introduction

- Feature selection
  - identifies a minimum set of relevant features
  - is applied before a learning algorithm
  - reduces computation costs
  - increases the speed up of learning process
  - increases the model interpretability
  - improves the classification accuracy performance

Introduction

- Feature selection
  - identifies a minimum set of relevant features
  - is applied before a learning algorithm
  - reduces computation costs
  - increases the speed up of learning process
  - increases the model interpretability
  - improves the classification accuracy performance
- Problem
  - finding the optimal number of genes for the feature selection
  - finding the optimal trade off between information loss (pruning excessively) and noise increase (pruning is too weak)

Goals

- Contribution
  - novel representation of genes as strings of bits
  - method which automatically selects the minimum number of genes to:
    - reach a good classification accuracy on the training set
    - improve accuracy of classifier model

Goals

- Contribution
  - novel representation of genes as strings of bits
  - method which automatically selects the minimum number of genes to:
    - reach a good classification accuracy on the training set
    - improve accuracy of classifier model
- Approach
  - our method first eliminates redundant features
    - genes do not add further information for classification
  - it exploits a set covering algorithm

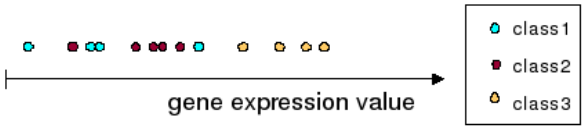
Approach

Gene representation

- Definition of the expression intervals of classes for each gene
  - let be K the number of classes
  - we define K intervals where
    - each interval contains the whole expression values for k-th class

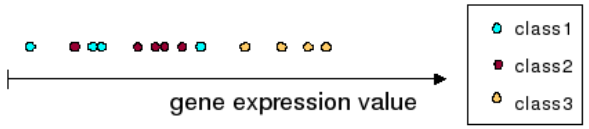
Gene representation

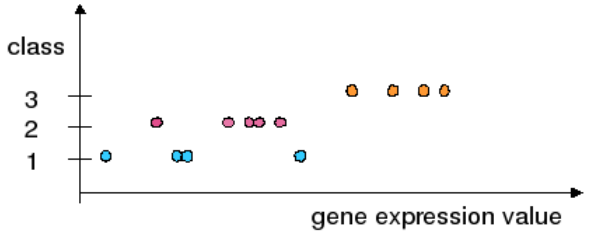
- Definition of the expression intervals of classes for each gene
  - let be K the number of classes
  - we define K intervals where
    - each interval contains the whole expression values for k-th class



Gene representation

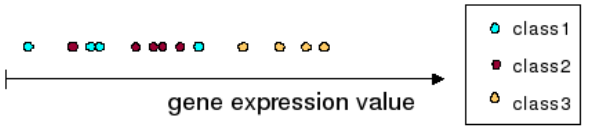
- Definition of the expression intervals of classes for each gene
  - let be K the number of classes
  - we define K intervals where
    - each interval contains the whole expression values for k-th class

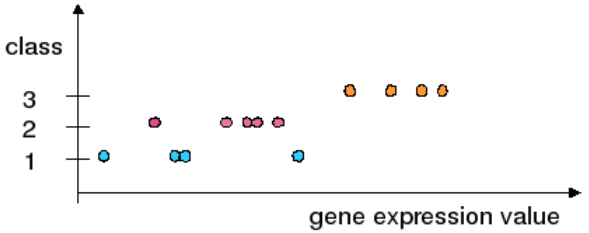


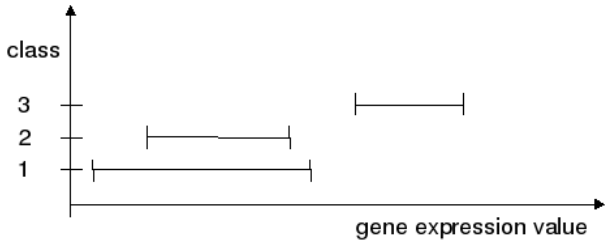


Gene representation

- Definition of the expression intervals of classes for each gene
  - let be K the number of classes
  - we define K intervals where
    - each interval contains the whole expression values for k-th class





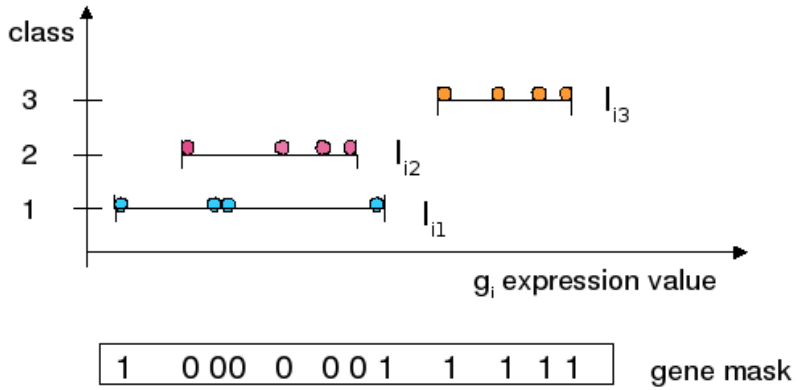


Gene Mask

- For each gene we define a gene mask
  - consists of an ordered sequence of M bits
  - M is the number of samples
  - an element of the gene mask is set to 1 if and only if the expression value of that sample only belongs to one class interval

Gene Mask

- For each gene we define a gene mask
  - consists of an ordered sequence of M bits
  - M is the number of samples
  - an element of the gene mask is set to 1 if and only if the expression value of that sample only belongs to one class interval



Mask covering algorithm (1)

Sample reduction Each sample which contains all 0 or 1 over the N gene masks is removed

- it is uninformative for the searching procedure

Gene reduction Each gene whose gene mask is a subsequence of another gene mask is removed

- if two or more genes have the same gene mask, the one with the largest variance in the expression values is selected

Reduced matrix evaluation The reduced matrix is evaluated by an optimization procedure which searches the minimum set of rows necessary to cover the binary matrix

- it is a min-max problem, it can be converted to a linear programming problem

Mask covering algorithm (2)

- A set covering algorithm is applied to the gene mask matrix
  - select the minimum set of genes whose ex-or generates a global mask of all ones
    - each sample is correctly classified by at least one gene

$$min \quad \sum_{i=1}^{N} g_i$$

$$\sum_{i=1}^{N} mask_{ij} \cdot g_i \geq 1, j = 1, ..., M$$

$$g_i \in \{0, 1\}$$

Mask covering algorithm (2)

- A set covering algorithm is applied to the gene mask matrix
  - select the minimum set of genes whose ex-or generates a global mask of all ones
    - each sample is correctly classified by at least one gene

$$min \quad \sum_{i=1}^{N} g_i$$

$$\sum_{i=1}^{N} mask_{ij} \cdot g_i \geq 1, j = 1, ..., M$$

$$g_i \in \{0, 1\}$$

- g_i is the i-th gene

Mask covering algorithm (2)

- A set covering algorithm is applied to the gene mask matrix
  - select the minimum set of genes whose ex-or generates a global mask of all ones
    - each sample is correctly classified by at least one gene

$$min \quad \sum_{i=1}^{N} g_i$$

$$\sum_{i=1}^{N} mask_{ij} \cdot g_i \geq 1, j = 1, ..., M$$

$$g_i \in \{0, 1\}$$

- g_i is the i-th gene
- $mask_{ij}$ is the j-th bit of i-th mask

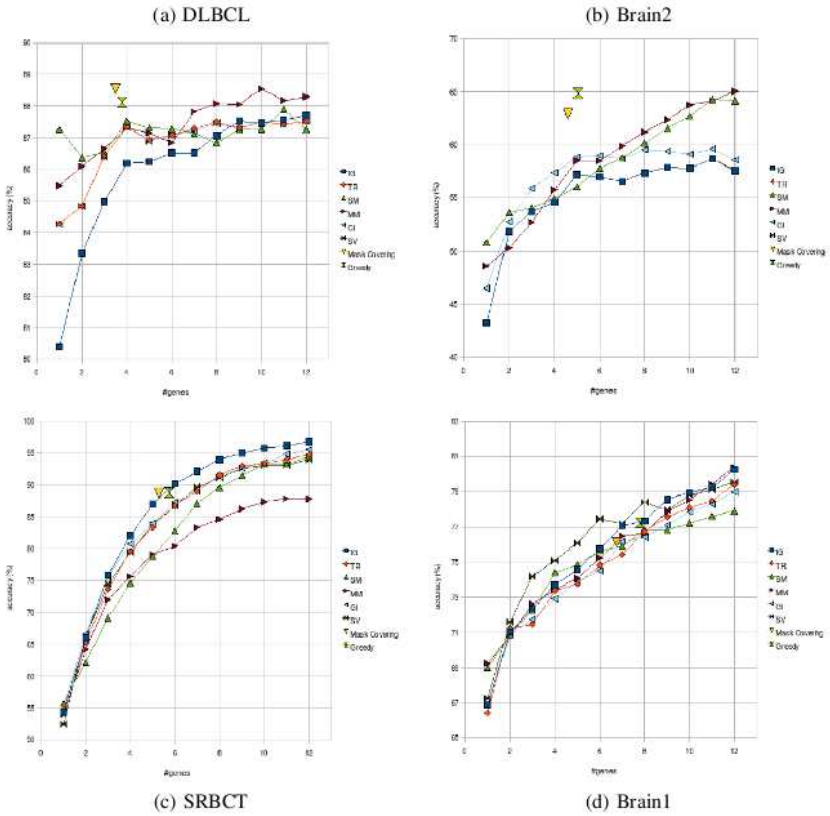
Experimental results

Experimental design

- Method compared:
  - Information Gain (IG)
  - Twoing Rule (TR)
  - Sum Minority (SM)
  - Max Minority (MM)
  - Gini Index (GI)
  - Sum of Variance(SV)
- 50 repetitions
  - 4-fold cross validation
- SVM classification
- Greedy vs Mask covering
- Datasets

Datasets	Samples	Genes	Classes
Brain1	90	5921	5
Brain2	60	10364	4
SRBCT	83	2308	2
DLBCL	77	5469	2

Experimental results



- Reduction rate

Dataset	Rate	Mask	Greedy
Brain1	68%	6.76	7.80
Brain2	92%	4.62	5.05
SRBCT	71%	5.28	5.75
DLBCL	77%	3.50	3.79

- Student t-test on classification perfomance
  - p-value < 0.01 on Brain2, SRBCT, DLBCL
  - p-value < 0.05 on Brain1

Biological validation

DLBCL dataset

- Mask covering includes
  - T-cell chemoattractant SLC
  - DNA replication licensing factor CDC47 homolog
- Greedy includes
  - DNA replication licensing factor CDC47 homolog
  - Cancellous bone osteoblast mRNA for GS3955
  - Chloride channel (putative) 2163bp
- all relevant for DLBCL disease [a]

[a] Shipp, M. and al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning", Nature Medicine, 8(1), pp 68-74, 2002

Conclusion

Conclusion

- Our method automatically selects the minimum number of genes needed to reach a good classification accuracy
- It exploits a novel representation of the gene capability to distinguish among classes, based on a bit mask
- The minimum set of genes is obtained by applying a set covering algorithm to this representation
- Experimental results show that our method reaches a very good accuracy with a low number of genes
  - these few genes can be used for further biological investigations
- Future work

Thanks for the attention!