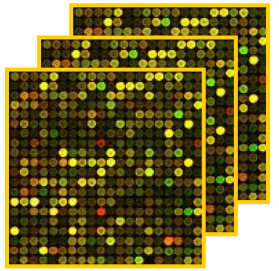
The Painter's Feature Selection for Gene Expression Data

Daniele Apiletti, Elena Baralis, Giulia Bruno, Alessandro Fiori

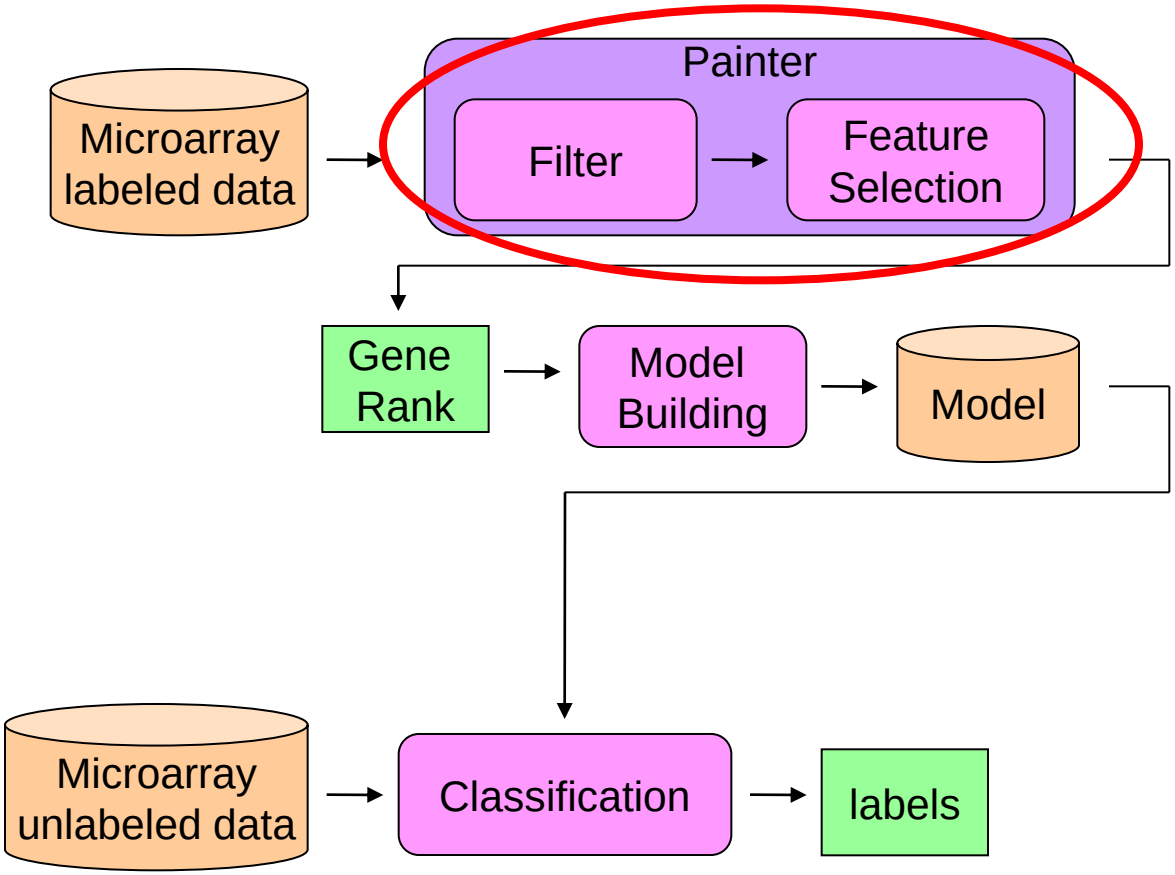
Lyon, 23-26 August 2007

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



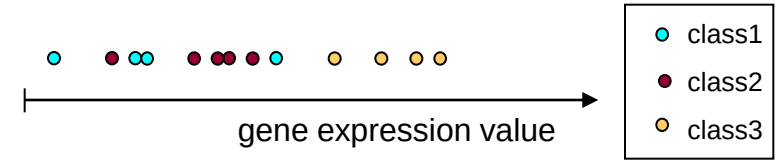
Framework

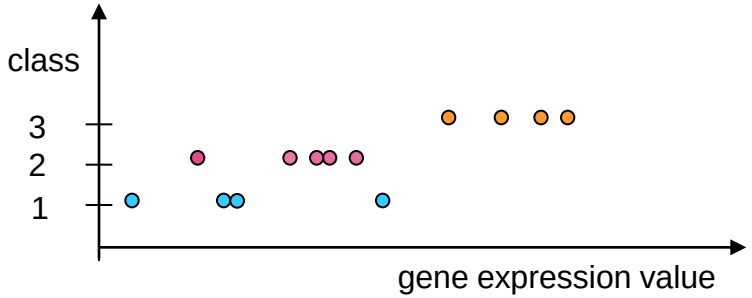

Painter
Filter
Feature Selection
Microarray labeled data
Gene Rank
Model Building
Model
Microarray unlabeled data
Classification
labels

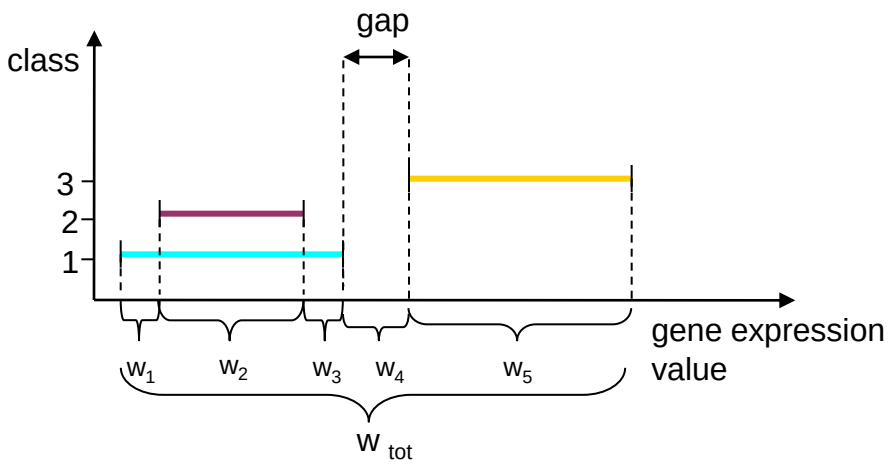
Feature selection

- From Painter's Algorithm in Computer Graphics to paint shadows
- Overlap score to measure:
  - common expression intervals in different classes
  - gaps between expression intervals among classes
- Bonus to genes with
  - largest gaps
  - few overlapping classes

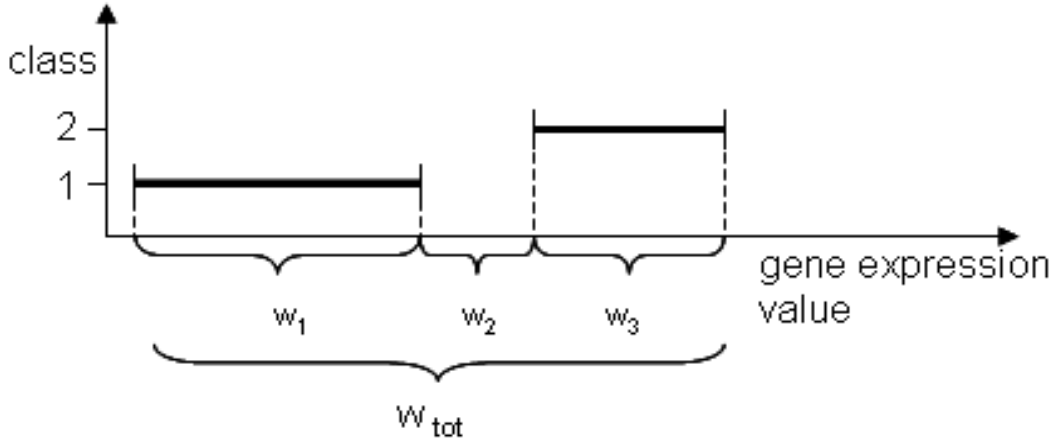
$$overlapscore = \frac{\sum_{i=1}^{n} c_i w_i}{w_{tot}}$$







Example

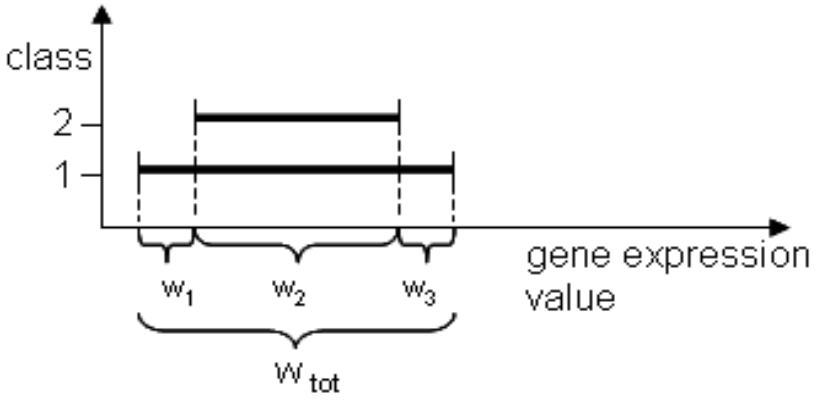


overlapscore =(1*6)+(1*4)/11=10/11

0 =< *overlapscore* <= 1 => no overlapping

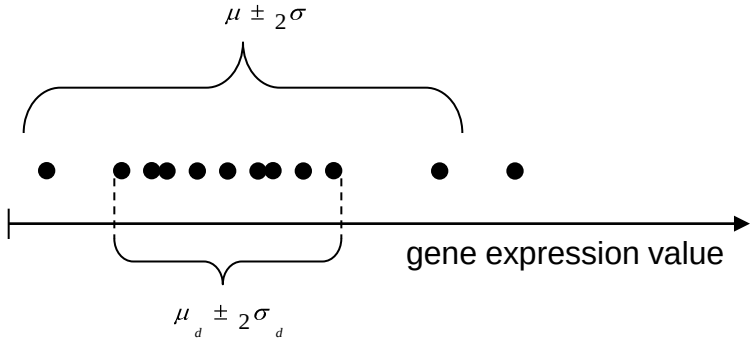
overlapscore= (1*1)+(2*4)+(1*1)/6=10/6

1 < *overlapscore* =< N_classes => overlapping



Filter

- Reason:
  - noisy data
  - outliers presence

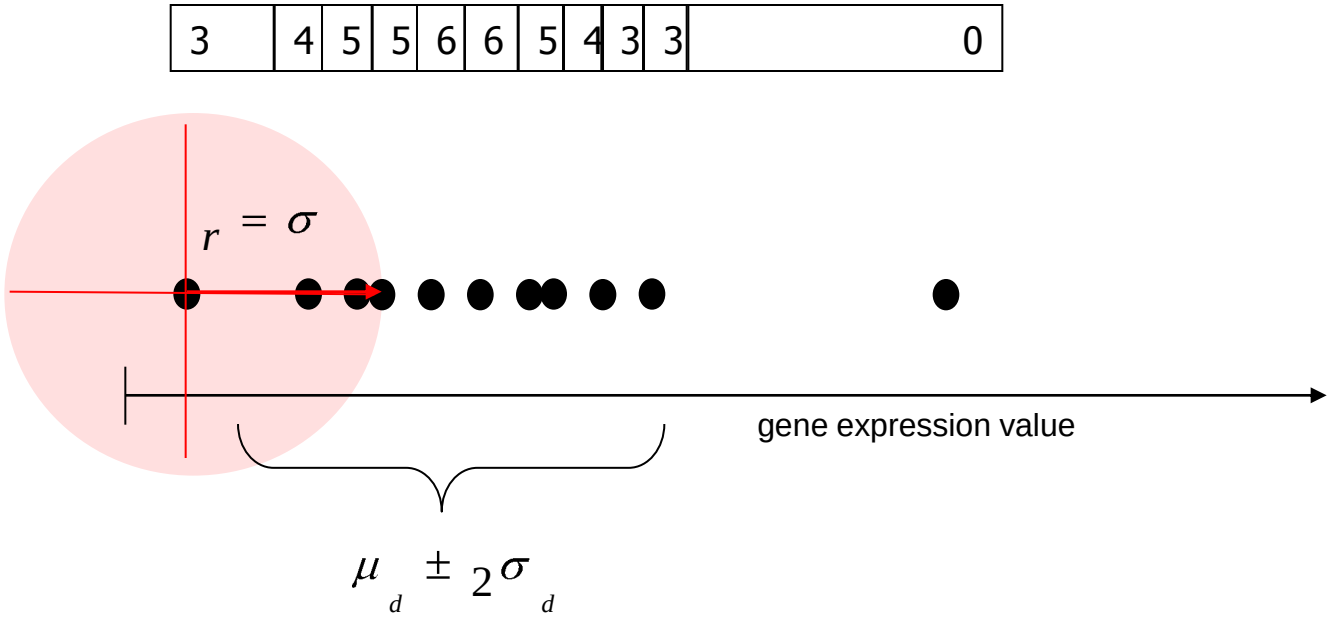


$$I_{ij} = \mu_d \pm c\,\sigma_d$$

$$\mu_d = \frac{1}{d_{tot}} \sum_{i=1}^{n} d_i e_i$$

$$\sigma_d = \sqrt{\frac{1}{d_{tot}} \sum_{i=1}^{n} d_i (e_i - \mu_d)^2}$$

Weighted interval



Experimental design

- 10-cross validation
- SVM kernel: Crammer and Singer (CS)
  - degree: 1
  - cost: 100
- Method compared:
  - analysis of variance (ANOVA)
  - signal-to-noise ratio in OVO (OVO)
  - signal-to-noise ratio in OVR (OVR)
  - ratio of variables between categories to within categories sum of squares (BW)

Datasets	Patients	Genes	Classes
Tumors9	60	5727	9
Brain1	90	5921	5
Brain2	60	10364	4

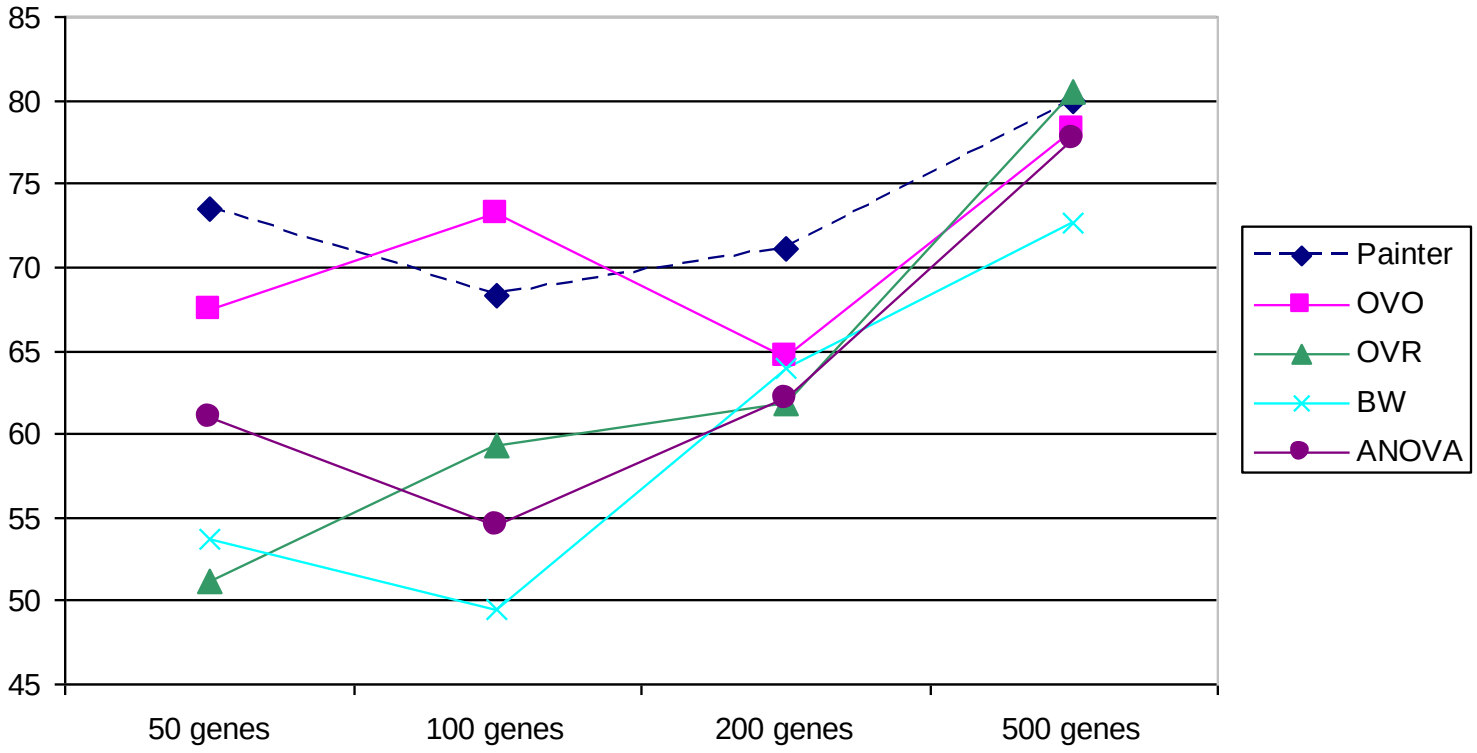
Experimental results (1)

- 200 genes selected

	Painter	OVO	BW	ANOVA	OVR
Brain1	86.7 8.8	**88.9 7.4**	86.7 9.1	**88.89 6.3**	87.8 7.0
Brain2	**70.7 20.0**	64.7 12.4	64.0 14.5	62.2 22.4	61,8 15.7
Tumors9	**71.0 24.4**	65.5 15.1	64.2 16.6	69.9 18.4	66.8 20.2
Average	**76.1 17.6**	73.0 11.6	71.6 13.4	73.7 15.7	72.1 14.3

Experimental results (2)

- trend on Brain2 dataset



Conclusion

- New method:
  - multi-class approach
  - based on new criterion of gene relevance
  - self adaptation to the datasets distribution
  - density based filter
  - smoothing outliers effect
- Results
  - robustness of the algorithm
  - can be applied to any dataset with continuous valued features
- Future work:
  - investigation of features groups with the same distiminating power
  - comparison with more feature selection techniques on other datasets

Thanks for the attention!