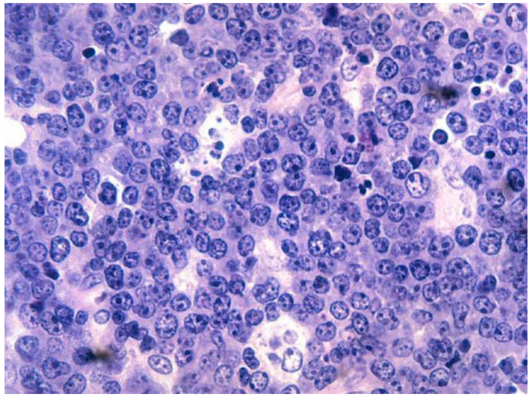
Prediction of tumor class from gene expression data using bagged decision trees



Ascension Vizinho-Coutry, avizinho@mathworks.fr

MathWorks France

Non-Clinical Statistics Conference | 2010

Agenda

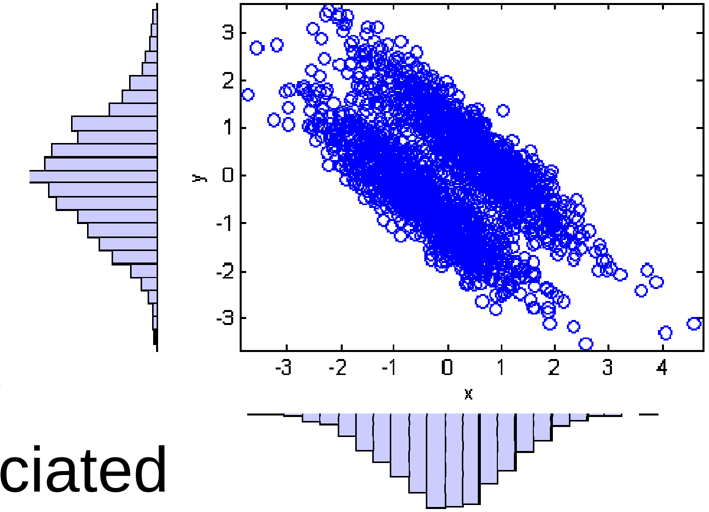
- Introduction:
  - Benefits of multivariate classification
  - Overview of multivariate classification capabilities in Statistics Toolbox™
- Tumor classification using gene expression data
  - Context (paper Khan *et al*.)
  - Bagged decision trees
  - Results
- Conclusion

Benefits of Multivariate Classification

- Classification methods help us to answer questions such as:
  - Are there subgroups in my dataset?
  - What are the similarities and differences between the groups?
  - Are the differences large enough that I can clearly discriminate between the groups, and predict what the group of a new observation is?
  - What variables explain the group differences?

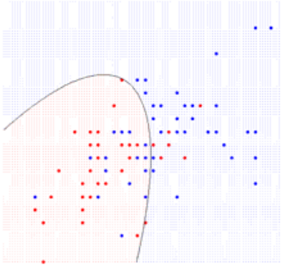
Benefits of Multivariate Classification

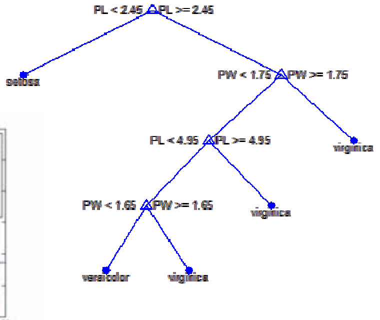
- Multivariate methods help to:
  - Simplify complex relationships between many variables
  - Find differences between groups that depend on the *correlations* between variables
  - Reduce the Type I errors associated with large numbers of univariate hypothesis tests

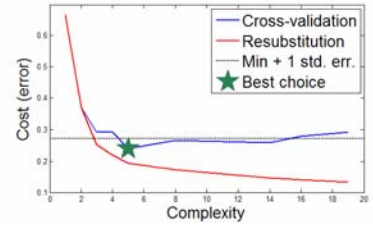


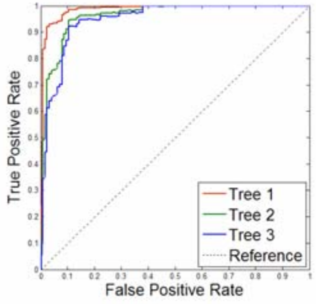
Multivariate Classification in Statistics Toolbox™

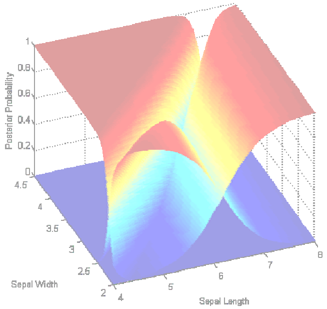
- Classification Algorithms
  - Linear and Quadratic Discriminant Analysis
  - Decision Trees
  - Naïve Bayes R2009a
  - Bagged Decision Trees R2009a R2009b
- Feature Selection Algorithms
  - Sequential Forward and Backward Feature Selection R2008a
- Model Selection and Evaluation Methods
  - Cross Validation R2008a R2008b R2009b
  - Confusion Matrices R2008b
  - ROC Performance Curves R2009a





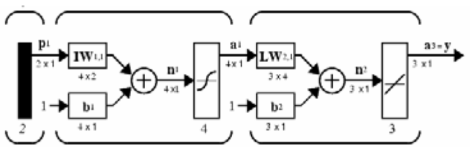


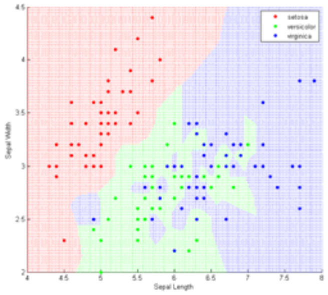


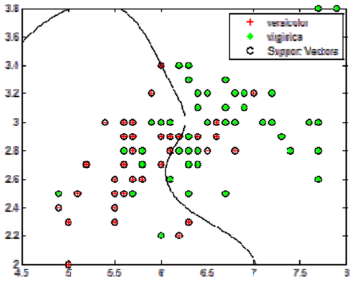


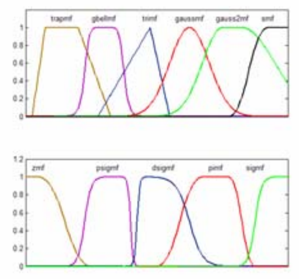
Related functionality in other toolboxes

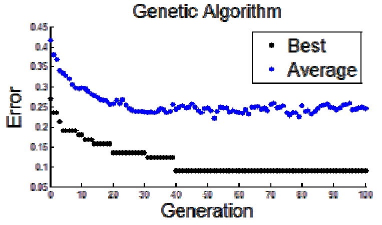
- Neural Networks Toolbox™
  - Neural network-based pattern recognition
- Bioinformatics Toolbox™
  - Support Vector Machines
  - K-nearest neighbours
  - Feature ranking, randomized feature selection
- Fuzzy Logic Toolbox™
  - Fuzzy classification methods
- Genetic Algorithms Toolbox™
  - GA-based feature selection





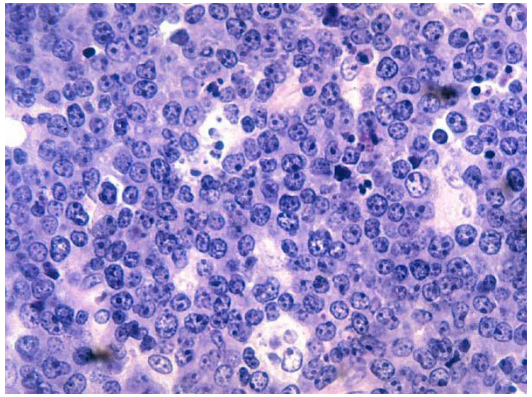






Tumor Classification using Gene Expression Data and Bagged Decision Trees

- Small, round blue-cell tumors (SRBCTs) belong to four distinct diagnostic categories:
  - neuroblastoma (NB),
  - rhabdomyosarcoma (RM),
  - Ewing family of tumors (EW), and
  - non-Hodgkin lymphoma, of which Burkitt lymphoma (BL) is a subset

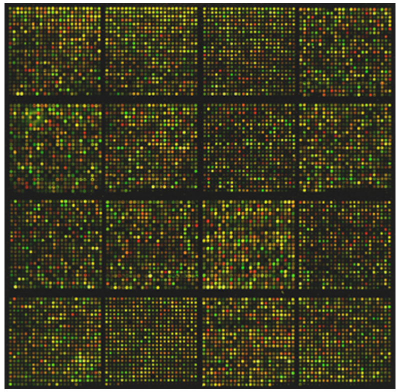


Malignant B-cell lymphocytes seen in Burkitt's lymphoma
Image credit: Louis M. Staudt, National Cancer Institute

- Accurate diagnosis is vital, leading to wide variation in treatment options and prognosis

- Difficult to distinguish by light microscopy
  - currently diagnosed by a combination of immunohistochemistry, cytogenetics, interphase fluorescence *in situ* hybridisation, and RT-PCR

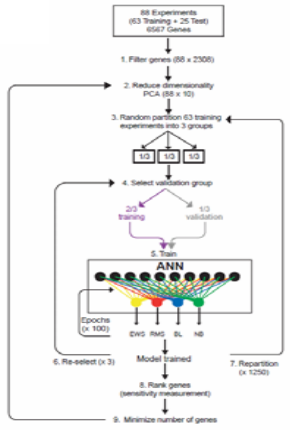
Tumor Classification using Gene Expression Data and Bagged Decision Trees



cDNA microarray, profiling expression levels of 8,700 genes
Image credit: Dr Jason Kang, National Cancer Institute

- Gene expression profiling with cDNA microarrays permits the measurement of multiple markers simultaneously
- But: gives rise to very large amounts of data, and requires multivariate methods to analyse

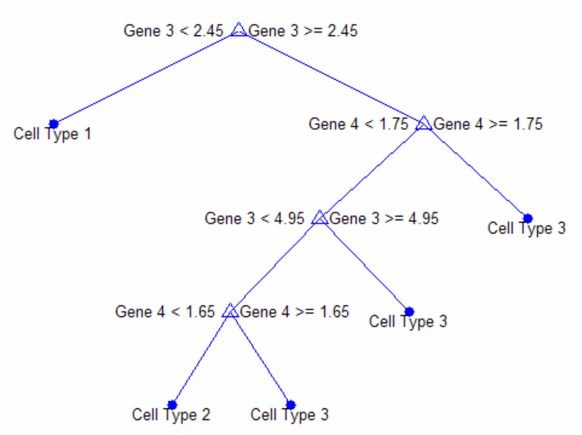
- Khan *et al.* used a neural network approach – here we use bagged decision trees



Khan J *et al.*, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7(6), 673-9, 2001

Bagged Decision Trees

- Bagged Decision Trees are constructed using two algorithms
  - An underlying decision tree classification algorithm
    - Takes a training dataset of pre-classified examples as input, and constructs a decision tree model that classifies future examples
  - A bagging (*b*ootstrap *ag*gregation) algorithm
    - Resamples the training dataset several times (bootstrapping), and builds a model from each; then aggregate these models together for a final classifier



```
Decision tree for classification
1  if Gene 3<2.45 then node 2 else node 3
2  class = Cell Type 1
3  if Gene 4<1.75 then node 4 else node 5
4  if Gene 3<4.95 then node 6 else node 7
5  class = Cell Type 3
6  if Gene 4<1.65 then node 8 else node 9
7  class = Cell Type 3
8  class = Cell Type 2
9  class = Cell Type 3
```

Breiman L, Random Forests. *Machine Learning* 45, 5-32, 2001

Data

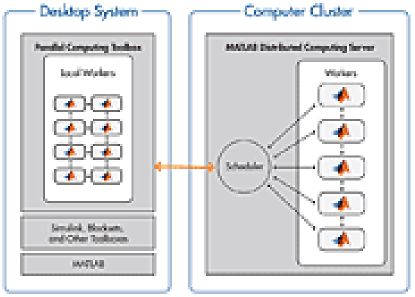
- 88 samples divided into a training set of 63 samples and a test set of 25 samples.
- Each sample is supplied with expression profiles of 2308 genes, collected using cDNA microarrays.
- The class of tumor to which the samples belong is also provided.

Approach based on the Tree Bagged Decision method

88 Samples
2308 Genes

Training set
65 samples

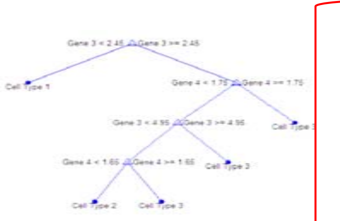
Test set
23 samples


Desktop System
Parallel Computing Toolbox
Local Workers
Simulink, Blocksets, and Other Toolboxes
MATLAB
Computer Cluster
MATLAB Distributed Computing Server
Scheduler
Workers

Pre-study 1
Tree Bagged
Decision Training
100 trees

Training 2
Top Variable
Selection (25)
500 trees for
Training

Test 3
Tree Bagged
Decision Test



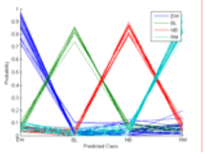
Evaluation

Classification SRBCT

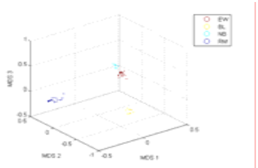
Confusion Matrix

	EW	BL	NB	RM
EW	23	0	0	0
BL	0	8	0	0
NB	0	0	12	0
RM	0	0	0	20

Confidence in the class prediction



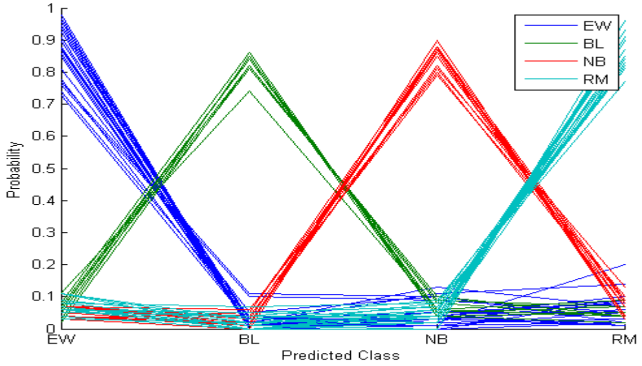
Proximity Matrix



1- Pre-study

- Take only 100 trees
- Compare actual classes to predicted classes

	EW	BL	NB	RM
EW	23	0	0	0
BL	0	8	0	0
NB	0	0	12	0
RM	0	0	0	20

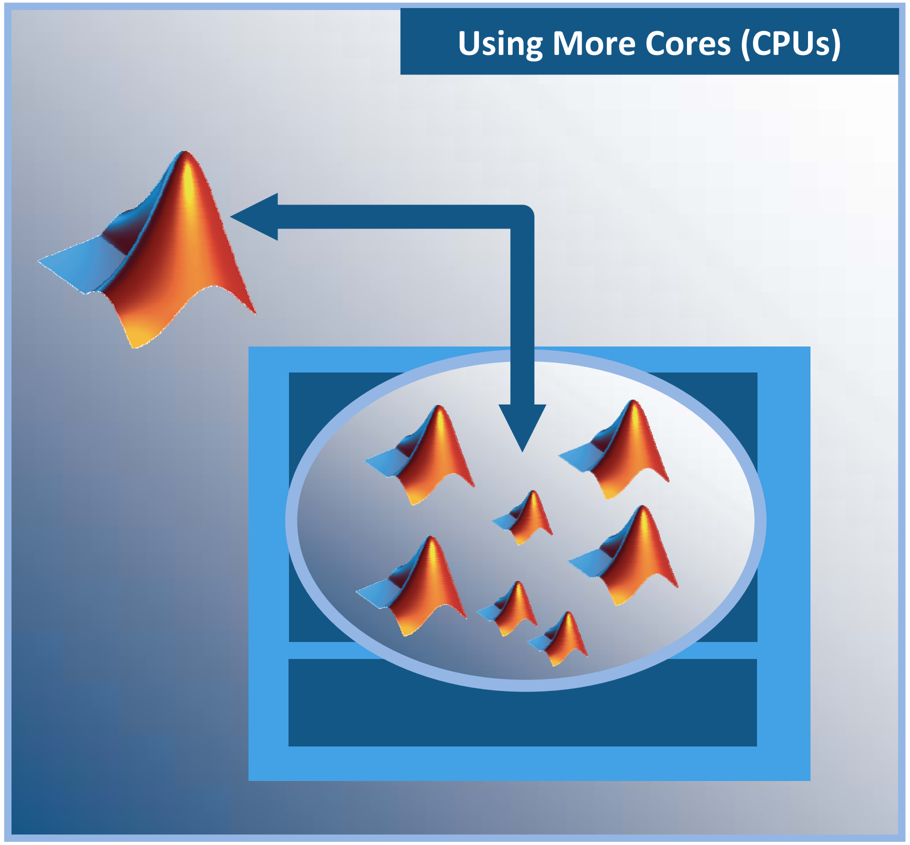


- Confidence on predicted classes
- Satisfactory results with out-of-bags samples (obtained from resampled dataset not used for model building)

Training

- Use Parallel Computing to select top variables because of permutation tests on out-of-bags variables

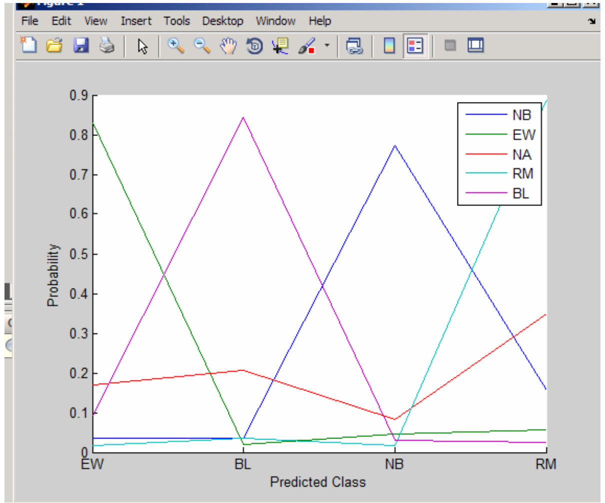
- Training using the 25 selected variables with 500 trees

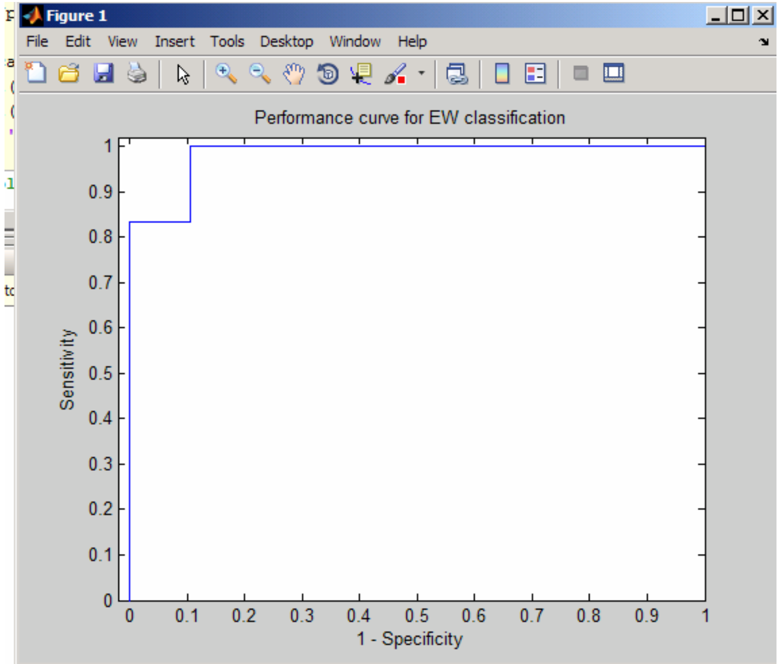


Evaluation

- Test using additional samples NA completely different

	NB	EW	NA	RM	BL
NB	6	0	0	0	0
EW	0	5	0	1	0
NA	1	1	0	3	0
RM	0	1	0	4	0
BL	0	0	0	0	3





Summary and Conclusion

- Bagged Decision Trees helped us to:
  - Build an accurate model to diagnose SRBCT class
  - Select a subset of important variables that could be taken forward into a custom assay
  - Provide class probabilities, not just hard classifications
- Confusion matrices, Performance Curves
  - Helped to evaluate models
- Integration with Parallel Computing Toolbox™
  - Accelerated computationally intensive learning tasks with large datasets

Thank you for listening

- For more information
  - Life Science applications:
    - http://www.mathworks.com/industries/biotech/
  - Statistics Toolbox™
    - http://www.mathworks.com/products/statistics/
  - Parallel Computing Toolbox™
    - http://www.mathworks.com/products/parallel-computing/