

A Comparison of Three Different Methods for Classification of Breast Cancer Data

D. Soria, J.M. Garibaldi, E. Biganzoli, I.O. Ellis

<http://www.cs.nott.ac.uk/~dqs>

ICMLA 2008, San Diego, US

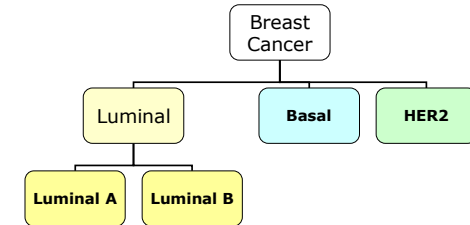
December 13, 2008

Outline

- Introduction and background
- Case study
- Automatic classification
- Results
- Ongoing work
- Conclusions

Background

- Identification of biologically **distinct groups** with clinical and prognostic relevance



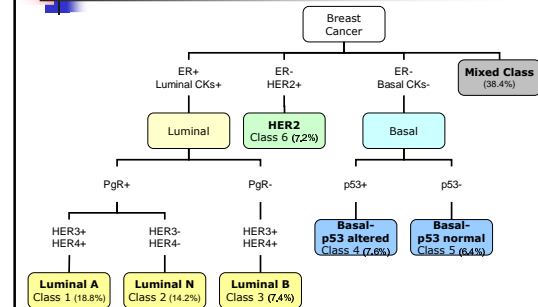
Background

- TMA technology allows **concomitant analyses** of many proteins on tumour samples
- Consensus between **clustering algorithms** identified **six core classes** (Garibaldi et al, 2008)
- **Not all patients** were classified

Case study

- Patients entered into the **Nottingham Tenovus Primary Breast Carcinoma Series** between 1986 and 1998
- **1076 cases** informative for all **25 biological markers**
- **Clinical information** (grade, size, age, survival, follow-up, etc.) available

Background



Classification techniques (1)

- Model-based classification for prediction of future cases
- Comparison of alternative classification techniques
- Aims
 - High quality prediction
 - Reduce number of biomarkers used
 - Prefer 'white-box' prediction model

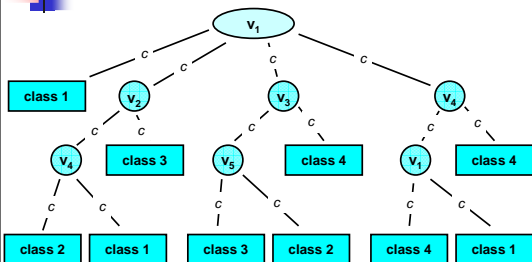
Classification techniques (2)

- Different classification techniques:
 - C4.5
 - Multi-Layer Perceptron Neural Network (MLP)
 - Naïve Bayes (NB)
- Software WEKA used
 - www.cs.waikato.ac.nz/~ml/weka

C4.5 classifier

- Generation of a decision tree
- Each attribute can be used to make a decision that splits the data into smaller subsets
- Information gain results from choosing an attribute for splitting the data
- Attribute with highest information gain is the one used to make the decision

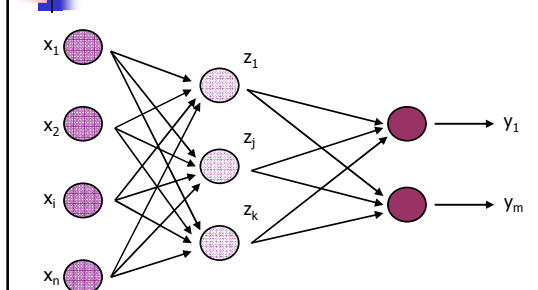
C4.5



Multi-Layer Perceptron

- Feed-forward artificial neural network
- Nonlinear activation function used by each neuron
- Layers of hidden nodes connected with every other node in the following layer
- Learning carried out through back-propagation

MLP



Naïve Bayes classifier

- Probabilistic classifier based on Bayes' theorem
- Good for multi-dimensional data
- Common assumptions:
 - Independence of variables
 - Normality

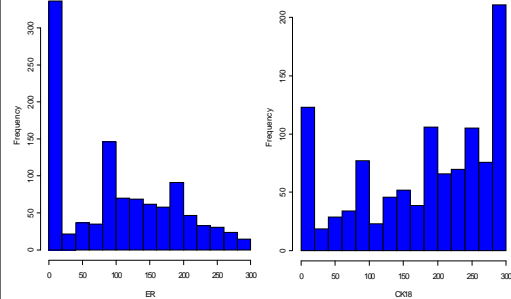
Running the algorithms

- 'In-class' subset (663/1076 patients)
- Each method run 10 times
- 10-fold cross validation
 - 9 folds for training, 1 fold for testing
- Accuracy evaluated (% of corrected classified instances) and mean returned

Results

	Method	Classified	Misclassified
Whole data	C4.5	582 (87.8%)	81 (12.2%)
	MLP	647 (97.6%)	16 (2.4%)
	NB	576 (86.9%)	87 (13.1%)
Ten markers	C4.5	581 (87.6%)	82 (12.4%)
	MLP	629 (94.9%)	34 (5.1%)
	NB	617 (93.1%)	46 (6.9%)

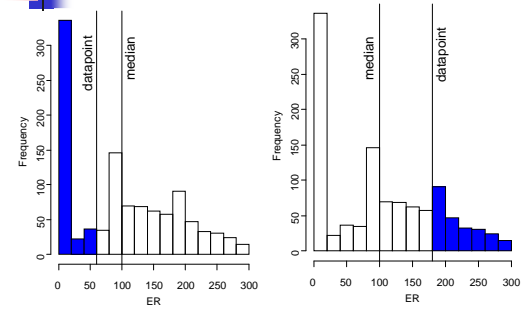
Variables' distribution



Ongoing work

- 'Non-parametric' version of Naïve Bayes classifier
- Based on ratio between areas under the histogram of each feature in each class

Ongoing work



Ongoing work

- `Non-parametric' version of Naïve Bayes classifier
- Based on **ratio between areas** under the histogram of each feature in each class
- **Improved** classification (10 markers):

•	<u>Classified</u>	<u>Misclassified</u>
•	637 (96.1%)	26 (3.9%)

- **Validation** on MLR datasets

Conclusions

- Model-based **classification** for high quality prediction
- **Best performance** from MLP
- NB improves when **features reduced**
- `Non-parametric' version of Naïve-Bayes method with **better** results

Thank you!

- Acknowledgments:



FP6 Marie-Curie EST Fellowship (FP6-007597)



- Contact: d.soria@cs.nott.ac.uk

References

1. D.M. Abd El-Rehim, et al. *High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses*, Int. Journal of Cancer, 116, 340-350, 2005.
2. L. Kaufman, P.J. Rousseeuw. *Finding groups in data*, Wiley series in probability and mathematical statistics, 1990.
3. A. Weingessel, et al. *An Examination Of Indexes For Determining The Number Of Clusters In Binary Data Sets*, Working Paper No.29, 1999.
4. G.A. Carpenter and S. Grossberg. *ART2: Stable self-organization of pattern recognition codes for analogue input patterns*. Applied Optics, 26, 4919-30, 1987.
5. J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, Los Altos, California, 1993.
6. S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2 edition, 1998.
7. G. John and P. Langley. *Estimating continuous distributions in bayesian classifiers*. Proceeding of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995.