**Definitions**

- **Biomarker** substance used as an indicator of a biologic state.
  - Its detection can indicate a particular disease state
- **IHC** (Immunohistochemistry) process of localising proteins in cells of tissue section; widely used:
  - in the diagnosis of cancer
  - to understand the distribution and localisation of biomarkers in different parts of a tissue

**Outline**

- Aims and objectives
- Definitions
- Framework
  - Clustering
  - Supervised learning
  - Validation on new data
- Conclusions and questions

**Aims and objectives**

- Develop an original framework with multiple steps to extract most representative classes from any dataset
  - Refine the phenotypic characterisation of breast cancer
- Move the medical decision making process from a single technique approach to a multi-technique one
  - Guide clinicians in the choice of the favourite and most powerful treatment

**Framework (1)**

```
Dataset
  Pre-processing
    Clustering
      HCA
      FCM
      KMI
      PAM
```

**Data pre-processing**

- Deletion of missing values
- Homogeneous variables
- Compute descriptive statistics
**Clustering**
- Four different algorithms:
  1. Hierarchical (HCA)
  2. Fuzzy c-means (FCM)
  3. K-means (KM)
  4. Partitioning Around Medoids (PAM)
- Methods run with the number of clusters varying from 2 to 20

**Definition of classes**
- Consensus clustering:
  - Align labels to have similar clusters named in the same way by different algorithms
  - Take into account patients assigned to the same group

**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

**If \( n \) is unknown...**
- Validity indices computation
- Defined considering the data dispersion within and between clusters
- According to decision rules, the best number of clusters may be selected

**Characterisation & agreement**
- Visualisation techniques
  - Biplots
  - Boxplots
- Indices for assessing agreement
  - Cohen’s kappa (\( \kappa \))
  - Rand

**Example: biplots (PCA)**
- K-Means
- PAM
Example: boxplots

Previous work: 4 groups

Our results: 6 groups

Framework (1)

Dataset
Pre-processing
Clustering

HCA  FCM  KM  PAM

n known?

Characterisation & agreement

Validity indices

Classes

Framework (2)

Classes

Supervised learning

C4.5  MLP - ANN  Naïve Bayes  NPBC

Supervised learning (1)

- Model-based classification for prediction of future cases
- Aims
  - High quality prediction
  - Reduce number of biomarkers used
  - Prefer ‘white-box’ prediction model
Supervised learning (2)

- Different classification techniques:
  - C4.5
  - Multi-Layer Perceptron Neural Network (MLP)
  - Naïve Bayes (NB) or Non-Parametric Bayesian Classifier (NPBC)

Characterisation of classes

- Biplots
- Boxplots
- Survival analysis
- Relation with other clinical information

Model based clustering

- Affinity Propagation (Frey and Dueck, 2007)
  - Both pair comparison and a probability model are used to choose the grouping
  - Real-valued messages are exchanged between data points until set of exemplars and corresponding clusters emerge
  - Suggestion of the number of clusters
  - CPU time for AP higher than KM
  - AP not included in the framework

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Classified</th>
<th>Misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>581 (87.6%)</td>
<td>82 (12.4%)</td>
</tr>
<tr>
<td>MLP</td>
<td>629 (94.9%)</td>
<td>34 (5.1%)</td>
</tr>
<tr>
<td>NB</td>
<td>617 (93.1%)</td>
<td>46 (6.9%)</td>
</tr>
<tr>
<td>NPBC</td>
<td>637 (96.1%)</td>
<td>26 (3.9%)</td>
</tr>
</tbody>
</table>

CPU time: KM v AP

- Graph showing comparison between KM and AP in terms of CPU time.
Validation of the framework
- Set of markers involved in breast cancer cell cycle regulation
- 347 patients and 4 markers
- Survival and grade available
- K-means and PAM used

Results
- \( k = 0.9 \) between KM and PAM
- 3 common classes found
  - Intermediate expression (class 1)
  - High expression (class 2)
  - Low expression (class 3)

Results: boxplots of classes

Results: Kaplan-Meier curves

Results: C4.5 decision tree

Main contributions
- Definition of six breast cancer classes
  - Two novel in literature
- Non-parametric approach for supervised learning
- Original framework for classes identification
Future work

- Clustering
  - Methods
  - Initialisation techniques
  - Distances
- Unclassified patients
- Complete analysis on histone markers

Thank You!

- Acknowledgements:
  - FP6 Marie-Curie EST Fellowship (FP6-007597)

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