Validation of a quantifier-based fuzzy classification system for breast cancer patients on external independent cohorts

Daniele Soria
Department of Computer Science
University of Westminster
Cavendish Campus
London, UK, W1W 6UW
Email: d.soria@westminster.ac.uk

Jonathan M. Garibaldi
School of Computer Science
University of Nottingham
Jubilee Campus
Nottingham, UK, NG8 1BB
Email: jon.garibaldi@nottingham.ac.uk

***Abstract*—Recent studies in breast cancer domains have identified seven distinct clinical phenotypes (groups) using immunohistochemical analysis and a variety of unsupervised learning techniques. Consensus among the clustering algorithms has been used to categorise patients into these specific groups, but often at the expenses of not classifying all patients. It is known that fuzzy methodologies can provide linguistic based classification rules to ease those from consensus clustering. The objective of this study is to present the validation of a recently developed extension of a fuzzy quantification subsethood-based algorithm on three sets of newly available breast cancer data. Results show that our algorithm is able to reproduce the seven biological classes previously identified, preserving their characterisation in terms of marker distributions and therefore their clinical meaning. Moreover, because our algorithm constitutes the fundamental basis of the newly developed Nottingham Prognostic Index Plus (NPI+), our findings demonstrate that this new medical decision making tool can help moving towards a more tailored care in breast cancer.**

***Index Terms*—Rule-based classification, Fuzzy rules, Validation, Breast cancer**

I. Introduction

Breast cancer is the most common cancer and cause of cancer death in women in the UK [1]. It is regarded as a heterogeneous group of diseases with complex and distinctive underlying molecular pathogenesis [2]. Computational methods have been developed to address this heterogeneity, to assist in predicting outcome, and to support clinical decision making in breast cancer management. In particular, clustering approaches have become more and more popular, and have been used on gene expression profiling and on patient tumour samples. We and others have applied protein biomarker panels, with known relevance to breast cancer, to large numbers of cases using tissue microarrays, exploring the existence and clinical significance of distinct breast cancer classes through clustering approaches [3] and consensus clustering methodologies [4]. This led to the identification of novel cancer subtypes [5], although a large proportion (38%) of patients presented mixed characteristics and remained unclassified. Subsequent studies [6], [7] have refined previous breast cancer classifications introducing a seventh biological group and reducing the number of patients 'not classified' in any particular class.

Alternative approaches may be used to 'soften' the rules of consensus clustering, like fuzzy rule-based systems (FRBS), which are the focus of our study. FRBS have been recently applied to classification problems in which non-fuzzy input instances need to be assigned to one of a given set of classes to produce high classification accuracy. Many approaches have been proposed for generating and learning fuzzy *if-then* rules from numerical data for classification problems [8].

The purpose of this paper is to present the validation of a previously developed data-driven subsethood-based fuzzy rule induction algorithm, named 'fuzzy quantification subsethood-based algorithm' (fuzzyQSBA) [7], [9] over three new independent sets of breast cancer data. The main intention of the proposed technique was to build a model that could be easily interpreted by non-experts in classification systems to refine previously identified breast cancer treatment groups [6]. Those seven breast cancer classes were derived using clinical expert knowledge, considering patient outcomes and response to treatments.

The structure of the paper is as follows: in section II, the methodology used and the data available are presented; the fuzzyQSBA algorithm is also quickly recalled and described. Results of the application of the algorithm to the new breast cancer data sets are reported in section III. Section IV concludes the paper with a discussion of the results, and suggestions for future work.

II. Materials and Methods

The dataset used for the development of the algorithm presented in [7] consisted of a cohort of 1,073 patients presented at Nottingham City Hospital between 1986 and 1998 with primary operable breast cancer. Among all the available information, the following ten markers were considered:

1) Estrogen Receptor (ER)
2) Progesterone Receptor (PgR)

3) cytokeratin 7/8 (CK7/8)
4) cytokeratin 5/6 (CK5/6)
5) EGFR
6) c-erbB2 (HER2)
7) c-erbB3 (HER3)
8) c-erbB4 (HER4)
9) p53
10) Mucin1 (MUC1)

This panel of protein biomarkers has been recently used to identify core classes which are clinically meaningful and well-characterised [6].

An additional set of 238 patients from the same Nottingham series for which immunohistochemistry data for the ten biomarkers was available is considered in this study as a first validation cohort for the fuzzyQSBA algorithm. The second independent external validation data set has been provided from Edinburgh, which comprised a cohort of 885 patients treated by breast conservation surgery, axillary node sampling or clearance, and whole breast radiotherapy between 1981–98 in Edinburgh (Edinburgh Breast Conservation Series) [10]. TMAs of the Edinburgh series were also stained for the same aforementioned biomarkers in Nottingham using the same procedures as previously described in [6], [11]. The third series of breast cancer biomarkers was collected in Hungary. This series comprised 368 screen detected and symptomatic consecutive cases diagnosed with primary breast cancer between 1999–2002 and operated at the Buda MÁV Hospital, Budapest, Hungary. A total of 271 cases were assembled in TMAs, the remaining cases were not included due to technical reasons or lack of relevant data [12].

Some missing values were present in the data sets considered, but only for specific markers. Records for ER, PgR, and HER2 were available for all patients in the three cohorts considered, making it possible to proceed with the validation of the fuzzyQSBA algorithm. The process in which missing values were dealt with is explained below when the fuzzyQSBA algorithm is recalled.

The same seven classes previously identified in [6] were considered, to be classified using the specified ten markers. The original distribution of patients in these seven groups is presented in Table I. It can be seen that 38 patients remained unclassified (either distant from all classes or presenting mixed characteristics).

TABLE I
ORIGINAL DATA CLASS DISTRIBUTIONS ('NC' MEANS NOT CLASSIFIED).

Class	Patients
1 (Luminal A)	288 (26.8%)
2 (Luminal N)	205 (19.1%)
3 (Luminal B)	186 (17.3%)
4 (Basal p53 altered)	113 (10.5%)
5 (Basal p53 normal)	96 (9.0%)
6 (HER2+/ER+)	62 (5.8%)
7 (HER2+/ER-)	85 (7.9%)
NC	38 (3.5%)
Total	1073

The fuzzyQSBA algorithm uses fuzzy subsethood measures, rule induction approaches, and fuzzy quantifiers to produce a list of linguistic rules which can then be used for classification purposes. The algorithm and its background theory are fully described in [7], together with a class assignment procedure used to obtain a 'hard' classification into groups. A schematic representation of the whole process is reported in Fig. 1.

The overall methodology requires different steps: firstly, fuzzy sets 'high' and 'low' are defined for each variable (biomarker) and the corresponding membership functions are computed. H-scores values for each biomarker in the training and testing data are therefore changed to two different values corresponding to the degree of membership to each set 'high' and 'low'. As mentioned before, during the validation phase, some missing records were present in the data sets used for testing. To deal with missing information, and to tackle uncertainty, the following procedure was used: a membership value equal to 1 was assigned to the missing entries for both the functions 'high' and 'low' rather than using sigmoid equations. By following the overall procedure so far, 'modified' training and testing data sets are produced.

Then, using some pre-specified t–norm and t–conorm, the fuzzy subsethood values (needed to generate the fuzzy rules) are computed from the 'modified' training data, and the fuzzy rules are also obtained. The number of rules created corresponds to the number of possible classification outcomes. In our case, seven breast cancer classes needed to be sought, so seven rules were created. To generate the fuzzy rules, quantifiers need to be placed before instances of a fuzzy set. An example of a quantified statement is "most students who get a high score are young", where '*most*' is the quantifier, while '*high*' and '*young*' are the fuzzy sets.

The last step of the algorithm concerns the application of the fuzzy rules to the 'modified' testing data to generate the membership values of every data point (patient) to each of the seven classes. A list of possibility values is then 'appended' to each patient to represent their degree of membership to each biological group. Possibilities are real numbers between 0 and 1, indicating the degree of membership of a specific data point to a particular class. To derive a 'hard' classification where each patient is assigned to only a particular group, the 'class assignment algorithm' as described in [7] is employed.

For the purpose of this paper the algorithm has been trained on the original Nottingham data [6], [7] and tested independently on the new Nottingham, Edinburgh and Hungary validation series. The whole algorithm and the validation process were coded using *R*, a free software environment for computing and graphics [13].

III. RESULTS

The fuzzyQSBA algorithm was applied to the new Nottingham validation cohort first, then to the Edinburgh series, and finally to the cohort from Hungary. The classification outcomes are reported in Table II, Table III and Table IV respectively.

It can be seen that, although the percentage of patients in each class is not always consistent with the original data, the

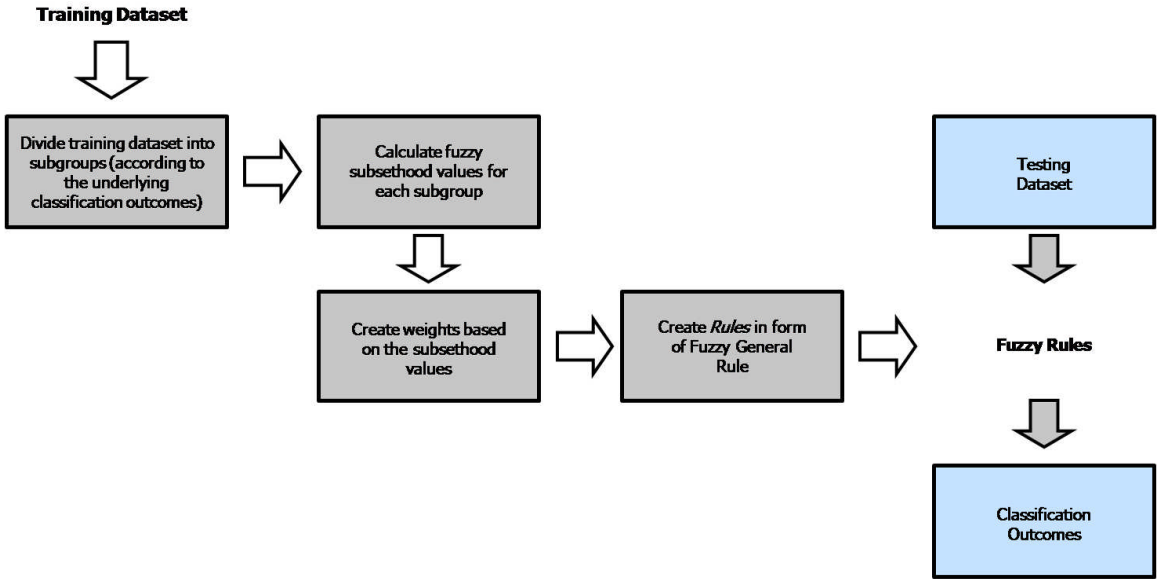

Fig. 1. Framework of fuzzyQSBA.

TABLE II
NEW NOTTINGHAM DATA CLASS DISTRIBUTIONS ('NC' MEANS NOT CLASSIFIED).

Class	Patients
1 (Luminal A)	68 (28.6%)
2 (Luminal N)	58 (24.4%)
3 (Luminal B)	39 (16.4%)
4 (Basal p53 altered)	17 (7.1%)
5 (Basal p53 normal)	28 (11.8%)
6 (HER2+/ER+)	8 (3.3%)
7 (HER2+/ER-)	8 (3.4%)
NC	12 (5.0%)
Total	238

TABLE III
EDINBURGH DATA CLASS DISTRIBUTIONS ('NC' MEANS NOT CLASSIFIED).

Class	Patients
1 (Luminal A)	241 (27.2%)
2 (Luminal N)	168 (19.0%)
3 (Luminal B)	157 (17.7%)
4 (Basal p53 altered)	102 (11.5%)
5 (Basal p53 normal)	79 (8.9%)
6 (HER2+/ER+)	32 (3.6%)
7 (HER2+/ER-)	55 (6.2%)
NC	51 (5.8%)
Total	885

TABLE IV
HUNGARY DATA CLASS DISTRIBUTIONS ('NC' MEANS NOT CLASSIFIED).

Class	Patients
1 (Luminal A)	84 (31.0%)
2 (Luminal N)	32 (11.8%)
3 (Luminal B)	65 (24.0%)
4 (Basal p53 altered)	18 (6.6%)
5 (Basal p53 normal)	36 (13.3%)
6 (HER2+/ER+)	15 (5.5%)
7 (HER2+/ER-)	16 (5.9%)
NC	5 (1.9%)
Total	271

seven classes in all three validation cohorts clearly retain their characterisation, as shown in Fig. 2, in Fig. 3, and in Fig. 4 compared with the original Nottingham data shown in Fig. 5.

Classes 1, 2 and 3 have a luminal profile, because they are characterised by high luminal CK7/8 and hormone receptor (ER and PgR) expression. Classes 1 and 3 (Luminal A and luminal B tumours) show high expression of CK7/8, ER, HER3 and HER4 but are separated by relatively lower levels of PgR expression in luminal B compared with luminal A tumours. In contrast, class 2 (luminal N tumours) shows differential expression of HER3 and HER4. Classes 4 and 5 (basal groups of tumour) are characterised by low luminal cytokeratin and high basal expression (CK5/6) along with showing a triple-negative phenotype (i.e., ER, PgR and HER2 are all under-expressed). They are, however, separated by p53 protein expression levels, resulting in the expression of either high p53 (basal-p53 altered) or low p53 (basal-p53 normal). The remaining classes (6 and 7) are characterised by high luminal cytokeratin and HER2 over-expression, resulting in a HER2+ profile. They are however distinguished in the expression of hormone receptors, with class 6 expressing ER (HER2+/ER+), and class 7 showing an ER- phenotype (HER2+/ER-). The unassigned patients show heterogenic expression of all ten markers.

IV. DISCUSSION

In this paper we presented the validation of a data-driven subsethood-based fuzzy rule induction algorithm,

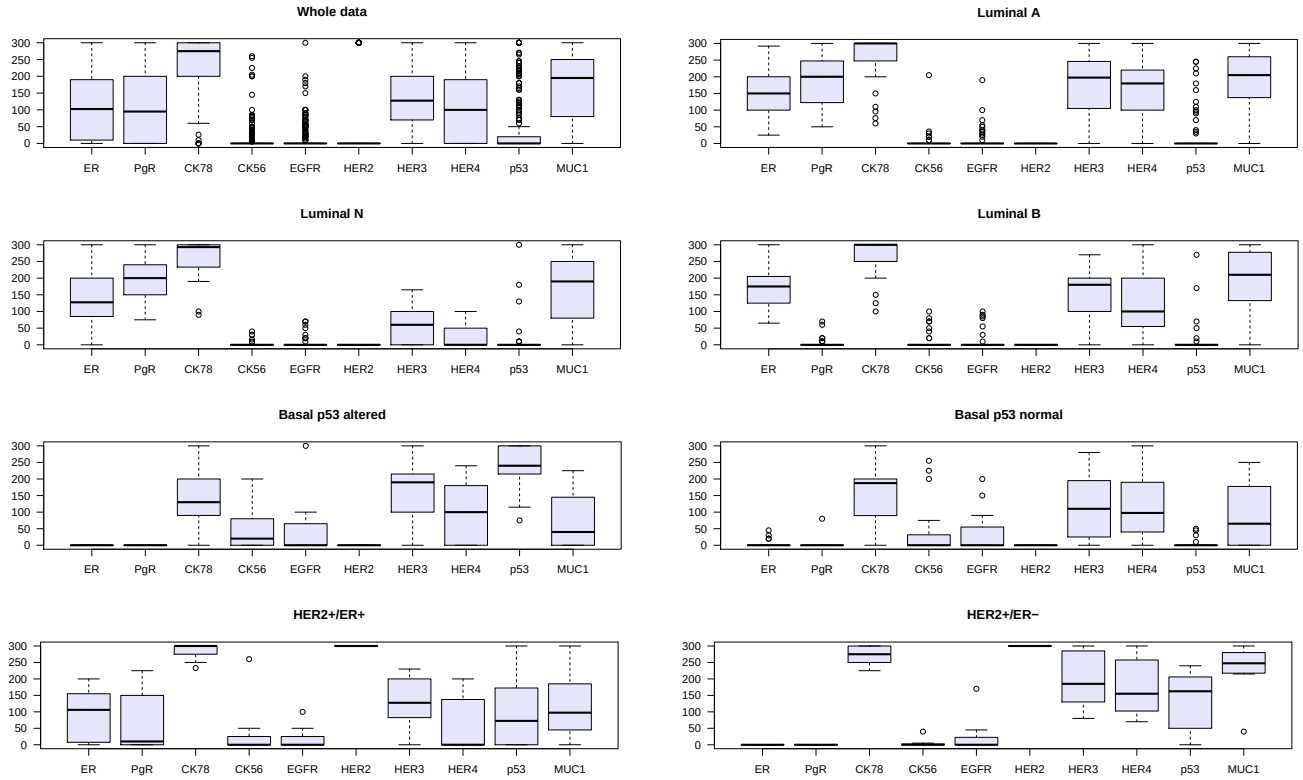

Fig. 2. Boxplots of ten markers (whole data and grouped by class) for the new Nottingham data (238 patients).

fuzzyQSBA [7], to new independent breast cancer data sets. Results show that the method is able to classify patients into the seven treatment groups previously identified [6] and demonstrate that the output classification meets the initial algorithm requirements for all the new validation series (Nottingham, Edinburgh, Hungary).

Several studies have been published in recent years investigating the usage of protein biomarker panels (with known relevance to breast cancer), in large numbers of cases using tissue microarrays, to highlight the existence and significance (in clinical terms) of distinct breast cancer classes [3], [5], [6]. However, although several cancer classes were clearly identified, many patients were left in a mixed-classified or unclassified group.

To ease the strict rules of hard-clustering, the use of fuzzy methodologies have become more and more important in addressing classification problems. In particular, FRBS have been employed to obtain high classification accuracy through linguistic rulesets. The fuzzyQSBA algorithm [8] is an example of such systems, which uses continuous fuzzy quantifiers to create the ruleset. This method, together with the class assignment algorithm presented in [7], has been used to obtain a refinement of the seven breast cancer classes introduced in [6]. In addition, the seven subclasses have significant differences in tumour characteristics and in clinical outcome, as reported in our most recent studies [14]–[16]. The results reported in this paper do not show the same percentage of patients (although somehow similar) in each class as in [7]; this can be due to the small size of the new Nottingham cohort and to some differences in the overall distribution of size, stage and grade of tumours for the Edinburgh and Hungary series which had an impact on the value of the biological markers. However, results for all validation cohorts do show how the distribution of the ten markers in the seven classes follow those previously presented and how the original requirements of the fuzzyQSBA algorithm (as reported in [7]) have been met for each validation cohort. An interpretation of the seven classes derived for the Edinburgh and Hungary series from a clinical perspective is out of the scope of this paper, but is presented in [15] and in [16] respectively.

In conclusion, we demonstrated that our methodology is capable to maintain the chracteristics of each of the seven breast cancer classes. Future work will focus on further validation of the algorithm on more external cohorts of patients from European and American centres; data for this additional validation are being currently acquired and gathered. We will also work on determining whether novel markers, such as Ki67, need to be incorporated in the model itself, as additional variables or to substitute some of the existing ones.

ACKNOWLEDGMENTS

We want to thank the Nottingham Breast Cancer Pathology group, the Edinburgh Cancer Research Centre, and the Second Department of Pathology, Semmelweis University, Budapest,

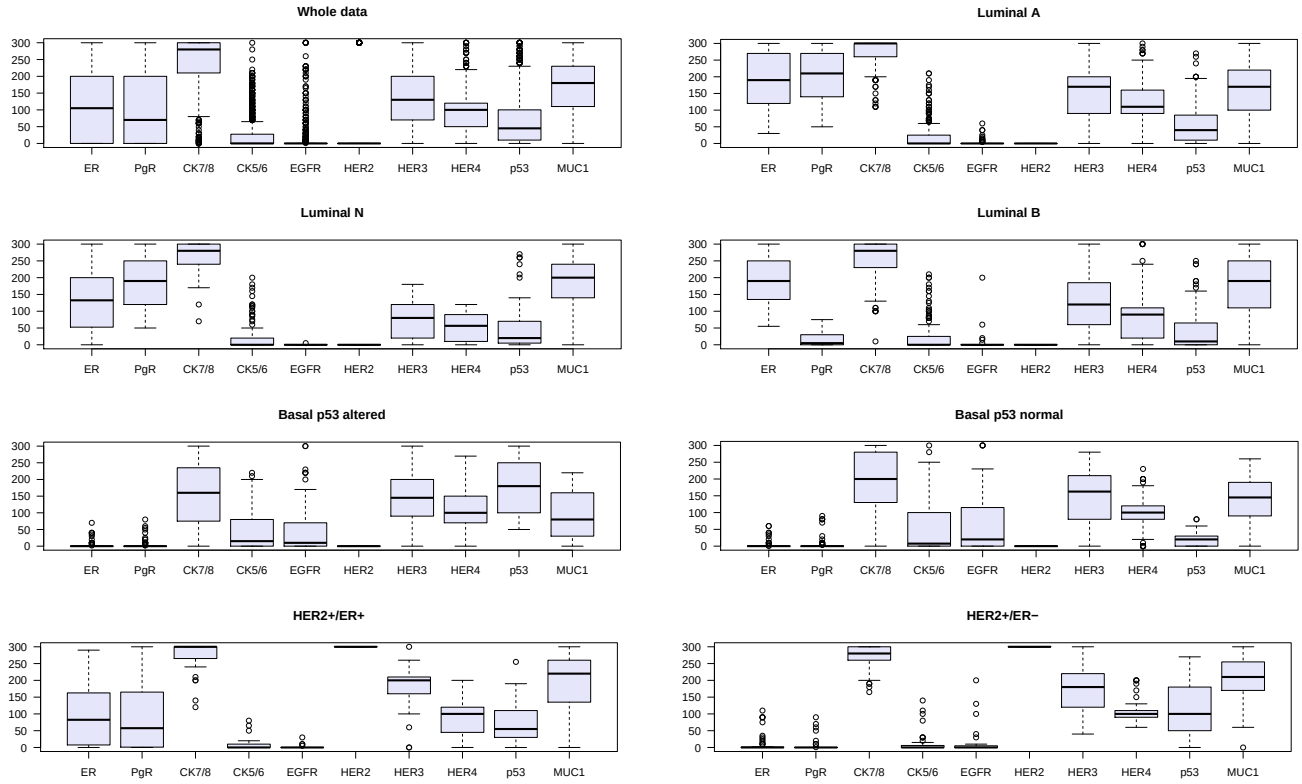

Fig. 3. Boxplots of ten markers (whole data and grouped by class) for the Edinburgh data (885 patients).

Hungary for collecting and providing the data and for helping with the clinical interpretation of the seven biological classes.

REFERENCES

[1] "Cancer Research UK: Cancer incidence for common cancers - UK statistics," 2012, accessed: 12 January 2016. [Online]. Available: http://info.cancerresearchuk.org/cancerstats/incidence/commoncancers

[2] I. Ellis, S. Pinder, A. Lee, and C. Elston, "A critical appraisal of existing classification systems of epithelial hyperplasia and in situ neoplasia of the breast with proposals for future methods of categorization: Where are we going?" *Semin Diagn Pathol*, vol. 16, pp. 202–208, 1999.

[3] F. Ambrogi, E. Biganzoli, P. Querzoli, S. Ferretti, P. Boracchi, S. Alberti, E. Marubini, and I. Nenci, "Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods," *Clinical Cancer Research*, vol. 12, no. 3, pp. 781–790, 2006.

[4] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam, "Consensus clustering and functional interpretation of gene-expression data," *Genome Biology*, vol. 5, no. 11, p. Article R94, 2004.

[5] D. Soria, J. Garibaldi, F. Ambrogi, A. Green, D. Powe, E. Rakha, R. Macmillan, R. Blamey, G. Ball, P. Lisboa, T. Etchells, P. Boracchi, E. Biganzoli, and I. Ellis, "A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients," *Computers in Biology and Medicine*, vol. 40, no. 3, pp. 318–330, 2010.

[6] A. Green, D. Powe, E. Rakha, D. Soria, C. Lemetre, C. Nolan, F. Barros, R. Macmillan, J. Garibaldi, G. Ball, and I. Ellis, "Identification of key clinical phenotypes of breast cancer using a minimised panel of protein biomarkers," *British Journal of Cancer*, vol. 109, no. 7, pp. 1886–1894, 2013.

[7] D. Soria, J. Garibaldi, A. Green, D. Powe, C. Nolan, C. Lemetre, G. Ball, and I. Ellis, "A quantifier-based fuzzy classification system for breast cancer patients," *Artificial Intelligence in Medicine*, vol. 58, no. 3, pp. 175–184, 2013.

[8] K. Rasmani and Q. Shen, "Subsethood-based fuzzy modelling and classification," in *Proceeding of the 2004 UK Workshop on Computational Intelligence*. Citeseer, 2004, pp. 181–188.

[9] ——, "Modifying weighted fuzzy subsethood-based rule models with fuzzy quantifiers," in *Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on*. IEEE, 2004, pp. 1679–1684.

[10] J. Thomas, G. Kerr, J. W.J., F. Campbell, L. McKay, H.-C. Pedersen, I. Kunkler, D. Cameron, U. Chetty, and J. Bartlett, "Histological grading of invasive breast carcinoma – a simplification of existing methods in a large conservation series with long-term follow-up," *Histopathology*, vol. 55, pp. 724–731, 2009.

[11] D. Abd El-Rehim, G. Ball, S. Pinder, E. Rakha, C. Paish, J. Robertson, D. Macmillan, R. Blamey, and I. Ellis, "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*, vol. 116, pp. 340–350, 2005.

[12] J. Kulka, A. Szász, Z. Németh, L. Madaras, Z. Schaff, I. Molnár, and A. Tőkés, "Expression of tight junction protein claudin-4 in basal-like breast carcinomas," *Pathology Oncology Research*, vol. 15, no. 1, pp. 59–64, 2009.

[13] J. Maindonald and W. Braun, *Data Analysis and Graphics Using R - An Example-Based Approach*. Cambridge University Press, 2003.

[14] E. Rakha, D. Soria, A. Green, C. Lemetre, D. Powe, C. Nolan, J. Garibaldi, G. Ball, and I. Ellis, "Nottingham prognostic index plus (NPI+): A modern clinical decision making tool in breast cancer," *British Journal of Cancer*, vol. 110, no. 7, pp. 1688–1697, 2014.

[15] A. Green, D. Soria, J. Stephen, D. Powe, C. Nolan, I. Kunkler, J. Thomas, G. Kerr, W. Jack, D. Cameron, T. Piper, G. Ball, J. Garibaldi, E. Rakha, J. Bartlett, and I. Ellis, "Nottingham prognostic index plus (NPI+): Validation of a clinical decision making tool in breast cancer in an independent series," *The Journal of Pathology: Clinical Research*, vol. 2, no. 1, pp. 32–40, 2016.

[16] A. Green, D. Soria, D. Powe, C. Nolan, M. Aleskandarany, M. Szász, A. Tőkés, G. Ball, J. Garibaldi, E. Rakha, J. Kulka, and I. Ellis, "Nottingham prognostic index plus (NPI+) predicts risk of distant metastases in primary breast cancer," *Breast Cancer Research and Treatment*, vol. 157, no. 1, pp. 65–75, 2016.

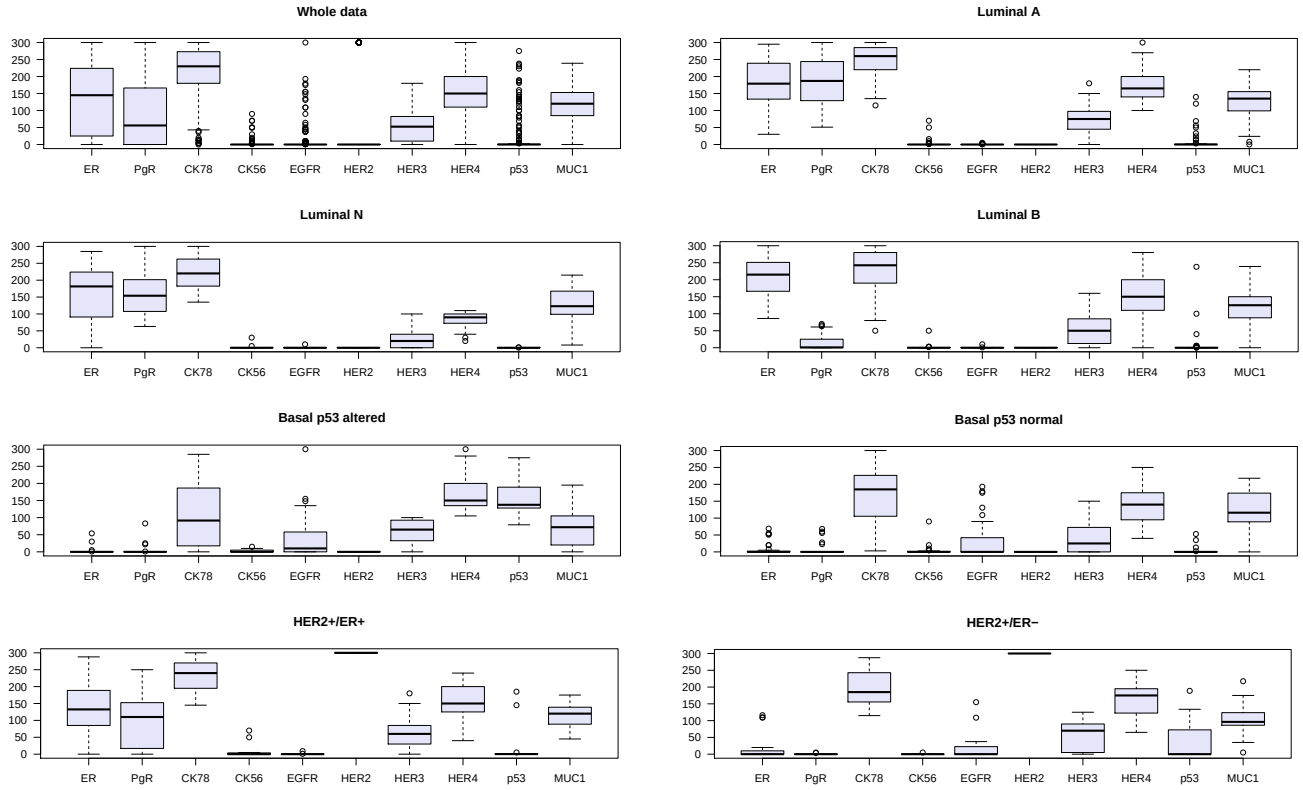

Fig. 4. Boxplots of ten markers (whole data and grouped by class) for the Hungary data (271 patients).

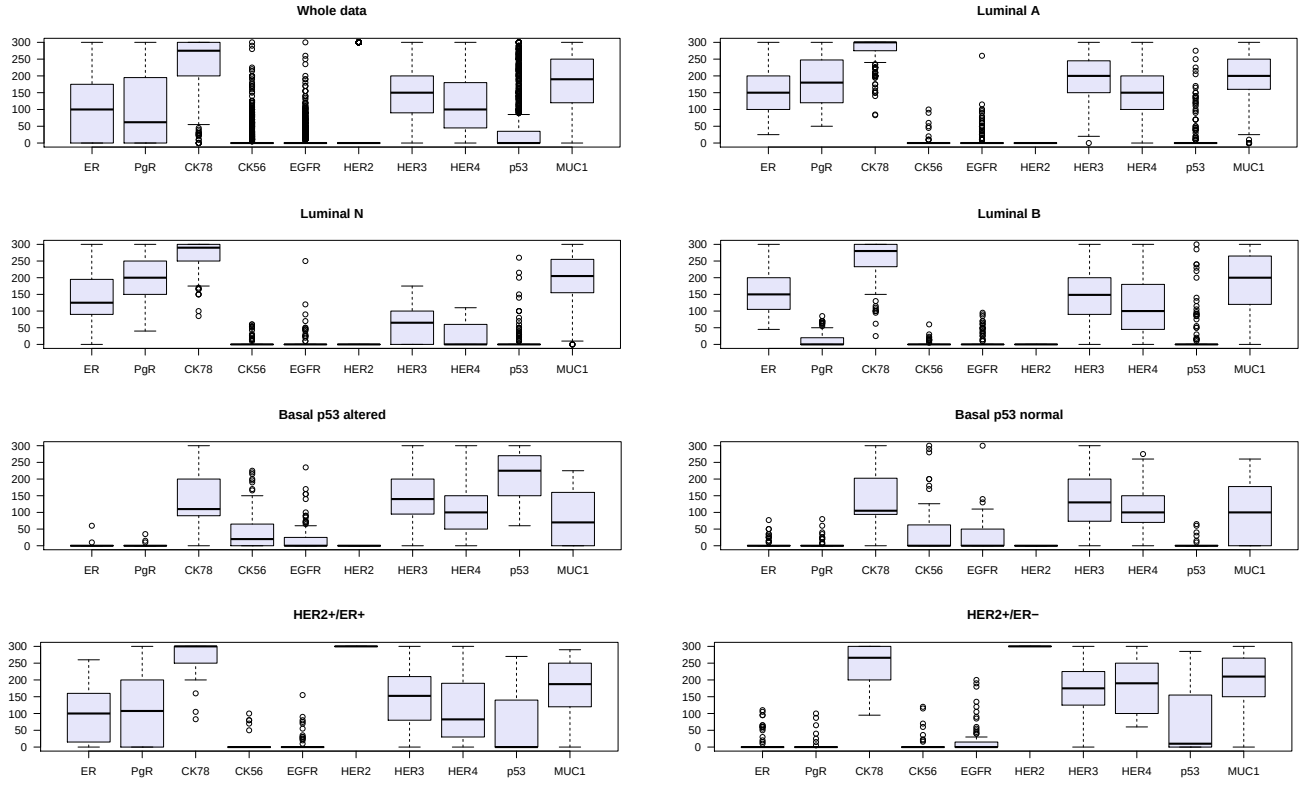

Fig. 5. Boxplots of ten markers (whole data and grouped by class) for the original Nottingham data (1073 patients).