A quantifier-based fuzzy classification system for breast cancer patients

Daniele Soria[a,*], Jonathan M. Garibaldi[a], Andrew R. Green[b], Desmond G. Powe[c], Christopher C. Nolan[b], Christophe Lemetre[d], Graham R. Ball[d], Ian O. Ellis[b]

[a]*School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK*
[b]*Department of Pathology, School of Molecular Medical Sciences, University of Nottingham, Queens Medical Centre, Derby Road, Nottingham NG7 2UH, UK*
[c]*Cellular Pathology, Nottingham University Hospitals NHS Trust, Derby Road, Nottingham NG7 2UH, UK*
[d]*School of Biomedical and Natural Sciences, The John van Geest Cancer Centre, Nottingham Trent University, Clifton Campus, Clifton Lane, Nottingham NG11 8NS, UK*

Abstract

Objectives: Recent studies of breast cancer data have identified seven distinct clinical phenotypes (groups) using immunohistochemical analysis and a range of different clustering techniques. Consensus between unsupervised classification algorithms has been successfully used to categorise patients into these specific groups, but often at the expenses of not classifying the whole set. It is known that fuzzy methodologies can provide linguistic based classification rules. The objective of this study was to investigate the use of fuzzy methodologies to create an easy to interpret set of classification rules, capable of placing the large majority of patients into one of the specified groups.

Methods and materials: In this paper, we extend a data-driven fuzzy rule-based system for classification purposes (called 'fuzzy quantification subsethood-based algorithm') and combine it with a novel class assignment procedure. The whole approach is then applied to a well characterised breast cancer dataset consisting of ten protein markers for over 1,000 patients to refine previously identified groups and to present clinicians with a linguistic ruleset. A range of statistical approaches were used to compare the obtained classes to previously obtained groupings and to assess the proportion of unclassified patients.

Results: A rule set was obtained from the algorithm which features one classification rule per class, using labels of *High*, *Low* or *Omit* for each biomarker, to determine the most appropriate class for each patient. When applied to the whole set of patients, the distribution of the obtained classes had an agreement of 0.9 when assessed using Kendall's Tau with the original reference class distribution. In doing so, only 38 patients out of 1,073 remain unclassified, representing a more clinically usable class assignment algorithm.

Conclusion: The fuzzy algorithm provides a simple to interpret, linguistic rule set which classifies over 95% of breast cancer patients into one of seven clinical groups.

Key words: Rule-based classification, Fuzzy rules, Linguistic ruleset, Breast cancer

1. Introduction

Breast cancer is the most common cancer and cause of cancer death in women in the UK [1]. It also leads in terms of numbers and complexity of available treatment options resulting in decision making difficulties regarding the most appropriate treatment choice [2]. Methods have been developed to assist in predicting outcome and to support clinical decision making in breast cancer management. One of the best known is the Nottingham prognostic index (NPI) [3], which is based on a combination of histopathological examination of tumour size, lymph node stage and tumour grading combined in a prognostic index formula [4]. The NPI is now used for the management of individual patients with breast cancer across Europe and elsewhere internationally. Recent data imply that breast cancer is a heterogeneous group of diseases with complex and distinctive underlying molecular pathogenesis [5]. However, the NPI does not contain sufficient information to represent and distinguish this heterogeneity. Further support for this hypothesis is provided by gene expression profiling which has identified distinct tumour groups that have direct clinical relevance in showing prognostic differences [6–8]. One of the major challenges in the computational analysis of such data is the curse of dimensionality because of the overwhelming number of variables measured (genes) versus the small number of samples [9]. In addition, due to experimental and technical reasons, there are large quantities of noise and redundancy in gene expression data, which may lead to building a prognosis predictor with poor performance [10].

To address the breast cancer disease heterogeneity, clustering approaches have become more and more popular, especially for discovering profiles in cancer with respect to high-throughput genomic data [11, 12]. Moreover, an alternative approach to gene expression profiling is to use established robust laboratory technology, such as immunocytochemistry on formalin fixed paraffin embedded 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 [13–16]. However, since different clustering algorithms result in different clusters, particularly when large multi-dimensional data sets are considered, consensus clustering methodologies have been used in recent studies [17–20]. In previous work [21], we applied different clustering algorithms and, through a consensus clustering approach, we identified novel cancer subtypes. This was done at the expense of not classifying a large proportion (38%) of patients.

Alternative approaches may be used to 'relax' the rules of consensus clustering such as rough sets or fuzzy classification methodologies. Rough set theory introduced by Pawlak in 1982 is a mathematical tool to deal with vagueness and uncertainty of information. This approach seems to be of fundamental importance to artificial intelligence, especially in the areas of machine learning and decision support systems [22]. Rough sets theory makes use of lower and upper approximations

*Corresponding author. School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK. Tel.: +44 115 84 68379. Email address: daniele.soria@nottingham.ac.uk

to set boundaries, and one of its main advantages is that it does not need any preliminary or additional information about data, such as grade of membership or the value of possibility in fuzzy set theory [23]. Parthaláin *et al.* have successfully used rough and fuzzy-rough set methods for the analysis of mammographic data [24]. However, Li and Wang [25] stated that the rules generated by rough sets are often unstable and have low classification accuracy. For this reason, and because we were not interested in eliminating redundant data (work on reducing the number of biomarkers had been previously undertaken [26]), we focused on fuzzy rule-based systems in our study.

Fuzzy rule-based modelling has become an active research field in recent years because of its unique merits in solving complex non-linear system identification and control problems. Primary advantages of this approach include the facility for the explicit knowledge representation in the form of *if-then* rules, a mechanism of reasoning in human understandable terms, the capacity of taking linguistic information from human experts and combining it with numerical information, and the ability of approximating complicated non-linear functions with simpler models. Unlike conventional modelling, where a single model is used to describe the global behaviour of a system, fuzzy rule-based modelling is essentially a multi-model approach in which individual rules (where each rule acts like a 'local model') are combined to describe the global behaviour of the system [27].

Fuzzy rule-based systems (FRBS) have often been applied to classification problems in which non-fuzzy input vectors are 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 [28, 29]. FRBS are used by Chang and Liu [30] for stock price prediction, while Ishibuchi and Yamamoto [31] show how the rule weight of each fuzzy rule can be specified in FRBS in the case of multiclass pattern classification problems. Of interest to this paper are data-driven FRBS for handling classification tasks.

There are many non-fuzzy classification algorithms currently available [32]. However, many of these classification algorithms may be very good in generalisation ability and so be very useful for classifying new instances, but lack comprehensibility of the generated models. In fact, most of the models generated by non-fuzzy classification algorithms contain numerical values and may not be linguistically interpretable. This makes it harder for the user to utilise the models for decision making purposes. Note that an automated-system, also known as a computer assisted system, is normally considered as a tool to assist experts or non-experts in decision making. Hence, interpretability of such a system should be regarded as highly important [33].

The purpose of this paper is to use a data-driven subsethood-based fuzzy rule induction algorithm, named 'fuzzy quantification subsethood-based algorithm' (fuzzyQSBA) [34] to refine previously identified breast cancer treatment groups [35]. In addition, using a rule simplification technique, a linguistic ruleset can be extracted from the algorithm. The main intention of the proposed technique is to build a model that can be easily interpreted by a non-expert in classification systems. The seven breast cancer classes presented by Green at al. were derived using clinical expert knowledge, considering patient outcomes and response to treatments. The under-

lying classification was firstly proposed by Soria et al. [21], where different clustering techniques were combined using a consensus approach and six clinically relevant groups were identified. The limitation of the six-classes approach reported by Soria et al. was the high number of patients who presented mixed class characteristics and therefore remained unclassified. From a clinical perspective, reducing the number of unclassified patients represents an important challenge in order to be able to advise them on the most accurate and effective treatment. Consequently, reducing the number of unclassified patients to a minimum was also a major objective.

The structure of the paper is as follows: in section 2, the background theory of fuzzy subsethood values, fuzzy quantifiers and subsethood-based fuzzy rule induction algorithms is reported and summarised. At the end of this section, the fuzzyQSBA algorithm is described. Section 3 describes the methodology used and the algorithm specifications. Results of the application of the algorithm to the breast cancer dataset are presented in section 4. Section 5 concludes the paper with a discussion of the results, and suggestions for future work.

2. Background theory

2.1. Fuzzy subsethood measures

Let A and B be two fuzzy sets defined on the universe U. The fuzzy subsethood value of A with regard to B, $S(B, A)$ represents the degree to which A is a subset of B:

$$S(B, A) = \frac{\sum_{x \in U} \nabla(\mu_B(x), \mu_A(x))}{\sum_{x \in U} \mu_B(x)} \tag{1}$$

where $S(B, A) \in [0, 1]$ and ∇ is a t-norm, such as the Łukasiewicz operator [36].

The above definition of fuzzy subsethood value can be extended to calculate the degree of subsethood for linguistic terms in an attribute value V to a decision class D. If $\{A_1, A_2, \ldots, A_n\} \in V$, it is possible to replace A with A_i and B with D in equation (1).

Many more subsethood measures have been developed and reported in literature [37]. However, in the rest of the paper we will use the definition reported in equation (1) as our goal is to extend the fuzzyQSBA algorithm [34].

2.2. Rule induction approaches

Fuzzy subsethood values have been used to promote certain linguistic terms as part of the antecedent of an emerging fuzzy rule. This approach involves three main steps [38]: a) classifying training data into subgroups according to the underlying classification results, b) calculating fuzzy subsethood values for every linguistic term, and c) creating rules based on fuzzy subsethood values.

The generation of fuzzy rules is therefore dependent on the fuzzy subsethood values between the decision to be made and the possible linguistic terms of the conditional attributes. In the approach proposed by Chen et al. [38], fuzzy rules are created subject to a pre-specified threshold value $\alpha \in [0, 1]$. Any linguistic term that has a subsethood value that is greater than or equal

to α will automatically be chosen as an antecedent for the resulting fuzzy rules. However, this methodology, termed the subsethood-based algorithm (SBA), assumes that all pieces of information gathered from the training data are equally important. This may not be the case in modelling many real problems.

For this reason, a weighted subsethood-based algorithm (WSBA) has been proposed [39], in which a certain weighting strategy has been taken to represent the degree of 'importance'. In particular, weights are created from the subsethood values to provide multiplication factors for each linguistic variable. They are calculated in an intermediate step (between steps b and c, mentioned above) using the following formula:

$$w(D, A_i) = \frac{S(D, A_i)}{\max_{j=1\ldots l} S(D, A_j)}, \qquad i = 1, \ldots, l$$

where $A_i \in \{A_1, \ldots, A_l\}$ is the i-th linguistic term of the linguistic variable A and D is the classification. The advantage of this method compared to the previous one is that it does not require any threshold value α. The crisp weights for each linguistic term can be considered as quantifiers.

A general case application of rule induction approaches is the well known 'Saturday morning problem' [40, 41], in which the weather on a Saturday morning (consisting of four attributes, each of which can take two or three linguistic values) is analysed to decide which sport is to be taken (classification result). Chen et al. [38], with their SBA method, achieved a better classification accuracy than the original subsethood-based algorithm [41]. When testing the WSBA on the same problem, Rasmani and Shen obtained even better results [39].

2.3. Fuzzy quantifiers

In general, a quantifier in logic can be expressed as $Q(x)A(x)$ where $Q(x)$ is a quantifier and $A(x)$ is a predicate for variable x. In classical logic, both the quantifier and the predicate can be represented by crisp sets. In fuzzy logic the quantifier may be applied to crisp or fuzzy sets. A quantifier based on fuzzy sets seems to be more suitable for quantifier based fuzzy models which are described in natural language.

Although different types of quantifier exist, the fuzzy relative quantifier Q will be considered here, in which $\mu_Q(q) \in [0, 1]$, with q defined on the real interval $[0, 1]$. In particular, Q possesses non-decreasing behaviour: $\forall q_1, q_2 \in Q, q_1 < q_2 \rightarrow \mu_Q(q_1) \leq \mu_Q(q_2)$. In general, the membership function $\mu_Q(q)$ of a quantifier Q has no direct meaning. Thus in evaluating a fuzzy quantified proposition, a quantification mechanism is needed to map the membership value $\mu_Q(q)$ such that:

$$F : (\mu_Q(q)) \rightarrow I \in [0, 1]$$

An example of a quantified statement is "most students who get a high score are young", where '*most*' is the quantifier, '*high*' and '*young*' are the fuzzy values A and B of equation (1) respectively. The result of evaluating the fuzzy relative quantifier is referred to as the truth-value of the quantifier, and is presented using notation T_Q [29].

The fuzzy quantification mechanism involves the definition of the existential quantifier, $\exists$, and of the universal quantifier, $\forall$. In addition to these, several different quantifiers can be defined, such as 'almost all', 'almost half', 'a few', etc. However, as small changes in the dataset might cause a change of the entire ruleset, a continuous fuzzy quantification method appears more appropriate.

Vila et al. [42] proposed a continuous fuzzy quantifier which uses linear interpolation between the two extreme cases of the existential quantifier $\exists$ and the universal quantifier $\forall$. In particular, the quantifier was defined as:

$$Q(A_{ij}, D_k) = (1 - \lambda_Q) \cdot T_{\forall, A/D} + \lambda_Q \cdot T_{\exists, A/D} \tag{2}$$

where Q is the quantifier for fuzzy set A relative to fuzzy set D and λ_Q is the degree of neighbourhood of the two extreme quantifiers. The truth value of the existential quantifier $T_{\exists, A/D}$ and the universal quantifier $T_{\forall, A/D}$ were defined as:

$$T_{\exists, A/D} = \Delta_{k=1}^{N} \mu(a_k) \nabla \mu(d_k) \tag{3}$$

$$T_{\forall, A/D} = \nabla_{k=1}^{N} (1 - \mu(d_k)) \Delta \mu(a_k) \tag{4}$$

where a_k and d_k are the membership functions of fuzzy sets A and D respectively, ∇ represents a *t-norm* and Δ represents the corresponding *t-conorm*. By using fuzzy subsethood values as the *degree of neighbourhood* (λ_Q) of the quantifiers, any possible quantifier that exists between the existential and universal quantifiers can be created in principle. Initially, all linguistic terms of each attribute are used to describe the antecedent of each rule. The reason for keeping this complete form is that every linguistic term may contain important information that should be taken into account.

2.4. FuzzyQSBA algorithm

The continuous fuzzy quantifiers are created using information extracted from data and behave as modifiers for each of the fuzzy terms. They can be then used to replace the crisp weights in WSBA, employing the quantification method proposed by Vila et al. [42]. Several reasons have been taken into account to support the use of Vila et al.'s approach:

i. The use of the degree of neighbourhood enables the implementation of continuous quantifiers. Thus, any possible quantifier can be created in principle.

ii. The relative quantifier based method proposed by Villa et al. can be adapted into WSBA easily, thanks to the structure of the WSBA general rule. Thus, the simplicity of WSBA can be preserved.

iii. Relative subsethood values can be used as the degree of neighbourhood of the fuzzy quantifiers. Thus, the two seemingly separate approaches are unified.

iv. This approach fulfills the desirable monotonicity and duality properties of quantification.

v. From a clinical point of view, continuous quantifiers are useful because their interpretability is normally regarded as highly important when developing decision support systems [43, 44].

The resulting new method is called fuzzyQSBA [34] and the induced ruleset can computationally be represented by

$$R_k = \nabla_{i=1\ldots m} \left(\Delta_{j=1\ldots n} \left(Q(A_{ij}, D_k) \nabla \mu_{A_{ij}}(x) \right) \right), k = 1, 2, \ldots, n \tag{5}$$

where $Q(A_{ij}, D_k)$ are fuzzy quantifiers as described in equation (2) and $\mu_{A_{ij}}(x)$ are fuzzy linguistic terms [33].

The crisp weights that were used in WSBA are herein replaced by fuzzy quantifiers. The main difference of fuzzyQSBA compared to WSBA is that in WSBA the weights for each linguistic term are crisp values and behave as multiplication factor for the linguistic terms. In fuzzyQSBA both the quantifiers and the linguistic terms are fuzzy sets. This offers flexibility as it enables the use of t-norm operators to interpret $\left(Q(A_{ij}, D_k) \nabla \mu_{A_{ij}}(x)\right)$ whilst guaranteeing that the inference results are fuzzy sets.

The use of fuzzy quantifier in QSBA also enables representation of the ruleset in a more natural way. This can be shown by the following example, in which a general rule is considered for the three different algorithms:

SBA "IF A is (A1 OR A2) AND B is (B2 OR B3) AND C is (C1) THEN Output is D".

WSBA "IF A is (A1 OR 0.09A2) AND B is (B2 OR 0.2B3) AND C is (C1) THEN Output is D".

fuzzyQSBA "IF A is ((almost all) A1 OR (a little)A2) and B is ((almost all)B2 OR (almost a quarter of)B3) AND C is ((almost all)C1) THEN Output is D".

Clearly, the use of fuzzy quantifiers make the model more readable, although the computation still needs to be performed using real numbers. Rules presented in the last example above are also useful for clinical judgment. For most of the biomarkers used in this work, there are no standard cut points used in clinical practice. The clinical cut point for ER and PgR, for example, is used to identify patients suitable for hormone therapy. However, there is evidence of a differential response to hormone therapy with increasing levels of these receptors supporting use of continuous data [45]. In addition, no evidence exists for a single clear HER2 status / protein level and response to treatment. For these reasons, the use of continuous rather than categorical data was deemed to be more appropriate for all markers, and hence the rules in the aforementioned form.

Based on the definitions of the fuzzy subsethood value, the existential quantifier and the universal quantifier (equations (1), (3) and (4)), it can be shown that if λ_Q is equal to zero then the truth-value of quantifier Q will also be equal to zero. Thus, during the rule generation process, the emerging ruleset is simplified as any linguistic terms whose quantifier has the truth-value of zero will be removed automatically from the fuzzy rule antecedents. Figure 1 shows the framework for this approach.

[Figure 1 about here.]

3. Methods

3.1. Algorithm specification

The dataset used for the development of the algorithm consists of a cohort of 1,073 patients presented at Nottingham city hospital between 1986 and 1998 with primary operable breast cancer [46]. Among all the available information, the following ten markers were considered:

1. Estrogen Receptor (ER)
2. Progesterone Receptor (PgR)
3. c-erbB2 (HER2)
4. cytokeratin 5/6 (CK5/6)
5. cytokeratin 7/8 (CK7/8)
6. EGFR
7. c-erbB3 (HER3)
8. c-erbB4 (HER4)
9. p53
10. Mucin1 (MUC1)

While 25 protein markers were originally used to derive the biological tumour classes presented by Soria et al. [21], this was subsequently reduced down to the above mentioned ten as the minimum number of markers compatible with retaining usefulness for clinical decision making [26]. The minimised panel of ten protein biomarkers has been recently used to identify core classes which are clinically meaningful and well-characterised [35]. Three of these classes had not been previously identified and, while their precise prognostic and therapeutic relevance is not yet clear, their elucidation serves as a basis for ongoing investigations in order to address these important factors. The core molecular classes identified by Green et al. are similar to those determined by gene expression profiling, but we have been able to refine the definition of the luminal and basal tumours into further distinct classes with different clinical outcome.

The same seven classes previously identified [35] were considered, to be classified using the specified ten markers. The original distribution of patients in these seven groups is presented in the first row of table 1. It can be seen that 76 patients remained unclassified (either distant from all classes or presenting mixed characteristics). In the development of the algorithm, all ten markers were used for the identification of the proper class. No missing values were present in the data set, so the set of 1,073 patients is complete with all information for the ten markers.

[Table 1 about here.]

The whole algorithm was coded using *R*, a free software environment for computing and graphics [47].

3.2. Class membership algorithm

The data-driven subsethood-based fuzzy rule induction algorithm, fuzzyQSBA [34] was used to determine the fuzzy class membership rules. In our particular case, the sets A and D of

equation (2) are the set of fuzzified data and the set of classification outcomes, respectively. In addition, it is important to note that the classification outcome D is not fuzzy. Thus, the value of d_k in equations (3) and (4) is always one.

Training and test data sets were transformed (fuzzified) using membership functions to create values representing the terms 'high' and 'low'. Membership functions were represented using sigmoid equations. In particular, for the term 'low' the function $f(x) = \frac{1}{1+e^{k(x-c)}}$ was used, while $f(x) = \frac{1}{1+e^{-k(x-c)}}$ was used for 'high'. In these equations, k represents a constant value defining the slope of the curve, while c is the fixed cut-off point for the specific variable.

For each variable, cut-off points c were selected to determine whether a particular value should be considered 'high' or 'low'. This was done by combining clinical knowledge and information extracted from the data. In particular, the median value of markers was used for ER, PgR, CK7/8, HER3, HER4 and MUC1. Clinical expertise was used for those markers (CK5/6 and HER2) for which clinical knowledge concerning the appropriate cut-off value is well-established, and for those which had a median equal to zero (EGFR and p53). An example of possible membership functions for the ten markers is shown in figure 2.

Having selected the cut-off c for each variable, the same values of c and the slope k was used for both 'low' and 'high' membership functions to maintain that $\mu(low) = 1 - \mu(high)$. Furthermore, although having different slopes of the membership functions for each different variable may better reflect the characteristics of markers, it was experimentally observed that using different slope values decreased the classification accuracy. Consequently, the same slope value was used for all the variables. This remains an issue open for further investigation.

[Figure 2 about here.]

The next step was to select the t-norms and t-conorms to be used for conjunction ('and') and disjunction ('or') operations. A t-norm is a kind of binary operation used in fuzzy logic which generalises conjunction in logic. T-norms are also used to construct the intersection of fuzzy sets. Different examples of t-norms have been proposed, with the most commonly used being the following:

- Minimum t-norm: $\mathrm{T}_{min}(a, b) = \min\{a, b\}$
- Product t-norm: $\mathrm{T}_{prod}(a, b) = a \cdot b$

T-conorms are dual to t-norms, generalising disjunction. Given a t-norm, the complementary conorm is defined by

$$\perp(a, b) = 1 - \mathrm{T}(1 - a, 1 - b)$$

Important t-conorms are those dual to prominent t-norms:

- Maximum t-conorm: $\perp_{max}(a, b) = \max\{a, b\}$

- Probabilistic sum: $\perp_{sum}(a, b) = a + b - a \cdot b$

In the development of the algorithm, the two different operator families (min-max and product-sum) were compared in both testing and training phases. It was found that the best overall performance was obtained when the min-max operators were used in the training phase (for deriving the classification rules) while product-sum were used in applying the classification rules, particularly in terms of the distribution of patients in the HER2 groups. This may be related to the fact that, with the min-max operators, there is a risk of losing some information. If, for instance, the minimum between two values has to be computed and one of them is always 0.01, then the result will not be affected by the second term being either 0.99 or 0.02. While we cannot explain the theoretical basis for this result, nevertheless, we selected the best overall model.

The goal of the class membership algorithm was to provide an indication of the likelihood of membership of each patient in each of the treatment classes. The output of the algorithm is a set of class membership values which may be interpreted as a set of *possibilities* of each patient belonging to each of the seven identified classes. Possibilities are real numbers ranging between 0 and 1, indicating the degree of membership of the particular patient to the particular class. This measure differs from a conventional probability, because in the latter case the sum of all probabilities of a single instance across all classes should be one. For possibilities, instead, every number should be between 0 and 1, but the sum across classes may be greater than 1. As a result, in the original dataset, seven extra columns were to be added by the algorithm for each patient. In each of these, a class membership (possibility) was reported. An example output is shown in table 2.

[Table 2 about here.]

3.3. Class assignment algorithm

It is important to distinguish the *fuzzy class membership* algorithm from the *class assignment* algorithm. The former takes the H-scores for the ten markers (from clinical measurement) and uses the fuzzy methodology described in section 3.2 to determine the fuzzy class membership of the patient in each of the seven classes. The latter subsequently takes the results obtained from the class membership algorithm, and uses a 'hard' strategy described below to determine the single 'best' class to represent each patient. This allows classes to be populated to allow comparisons with previous classifications and to meet algorithm specifications. The class assignment algorithm works as follows. Once a patient has been assigned a membership value for each class, the first and the second highest membership values are considered. If the difference between them is greater than a specified threshold, then the patient is assigned to the class with the maximum membership. If the difference is less than the threshold but the second maximum is in the same class family (luminal / basal / HER2) as the first maximum, then the patient is also assigned to the class with maximum membership. Otherwise the patient is assigned to the 'not classified' group. The specific

values of class assignment thresholds are not revealed in this paper as it is intellectual property of a spin-out company, Nottingham Prognostic Ltd [48], which is commercialising the decision support system.

3.4. Verification and Validation

Once completed, the algorithm was verified to assess whether it fulfils its requirements using the same 'internal' dataset. Following suggestions from clinicians, it was agreed that a suitable final classification should have between 12% and 15% of patients in HER2 classes (6 and 7 combined), while the number of 'not classified' patients should remain lower than 5%. Cohen's kappa index [49] and Kendall's tau coefficient [50] were used to measure the agreement between old and new classifications.

To avoid the over-fitting problem and issues about performing a test on self, the method underwent preliminary validation on novel data to determine whether it is applicable to other sources. An additional set of 238 patients, recently added to the Nottingham Tenovus Primary Breast Carcinoma Series [46] was used. Information about the ten biomarkers was available for all patients. As a first measure of comparison between obtained results, boxplots of the marker distributions in each class were created for the original (1,073) cases and the new (238) cases. Marker distributions were analysed in each class using Kruskal-Wallis tests.

A complete and thorough independent validation is still required in order to confirm the algorithm for clinical use. To perform this further validation, new data are currently being collected, and the whole validation process will be the subject of future work.

4. Results

The algorithm was run over the entire data. While the training of the algorithm was performed on the original dataset omitting the 76 not-classified cases (i.e. 997 cases), the whole dataset (1,073 cases) was used for testing purposes. Having defined all the necessary terms, equation (5) could be applied to define the ruleset and to compute membership values to each class.

4.1. Class membership algorithm

The linguistic rules table was generated using the quantifiers obtained by equation (2) and the cut-off points. In particular, the quantifiers table contained values for the 'high' and 'low' rules. The difference d between these two values was compared with a threshold λ and the terms *High*, *Low* and *Omit* were placed in the linguistic rules table using some rule simplification techniques. In particular, if the absolute value of d was lower than λ then *Omit* was entered in the table. If d was greater than zero, then *High* was entered, otherwise, if d was smaller than zero, *Low* was placed in the table. By using this procedure, the linguistic rules table reported in table 3 was obtained.

[Table 3 about here.]

In table 3, *Omit* means that the specified marker is not considered for the respective class membership. As all markers appear in at least one class rule, then in general all ten markers are needed for any new case. It might be possible, of course, to implement a 'step-wise' algorithm that measures the minimum number of markers that characterise any one single class (four markers for class 3) and assess whether they match. If so, the sample could be assigned to that class; if not, then more markers need to be measured. By doing so, it would be possible to reduce the number of markers measured for some samples, but at the expense of extra complexity. Note that the class assignment algorithm outlined in section 3.3 would also need detailed alteration, as the algorithm presented requires all seven class memberships to be calculated as input to the algorithm. We propose the simpler option of measuring all ten markers for all cases.

Table 3 was then compared to the ruleset defined by expert clinicians following the classification obtained by Green et al. [35] and reported in figure 3. It can be seen that ER, HER2, PgR and p53 are completely concordant, while CK 7/8 and CK 5/6 clearly identify the basal group (classes 4 and 5). HER4 discriminates between classes 1 and 2, while HER3 is also relevant in the characterisation of the latter. It is worth noting, in fact, that re-running a similar algorithm without considering the HER3 marker (i.e. using a 9-marker dataset) leaves a considerable number of patients originally classified in class 1 being assigned to class 2 (results not shown). The relevance of HER3 marker is open for future assessment. The new ruleset is shown in decision tree format in figure 4. By comparing the last two figures, it can be seen that the new ruleset considers fewer markers, especially for the luminal sub-groups.

[Figure 3 about here.]

[Figure 4 about here.]

4.2. Class assignment algorithm

By using the class assignment rules described above, the final classification was obtained as shown in table 1 (second row).

4.3. Verification and Validation

The HER2 group represented 13.7% of the total number of patients, while the 38 unclassified constituted 3.5%. These results are in agreement with the original algorithm specifications and requirements reported above. Moreover, clear agreement between the new classification and the original one was shown when calculating the Cohen's kappa and Kendall's tau coefficients; the former was equal to 0.72, while τ reached 0.89.

When the final model was applied to the new smaller dataset of 238 patients for testing, the seven classes were populated and contained 68, 58, 39, 17, 28, 8 and 8 patients, respectively. The remaining 12 patients (5%) presented mixed characteristics and were considered NCs. Boxplots of

the classes for the original data and the validation data are shown in figure 5 for comparison. It is clear that the newer groups have the same general characteristics of the original luminal, basal and HER2 groups and subgroups. These latest results are also confirmed by Kruskal-Wallis tests performed on single variables in different classes.

[Figure 5 about here.]

5. Discussion

This paper has presented a data-driven subsethood-based fuzzy rule induction algorithm, fuzzyQSBA [34] and its application to a breast cancer dataset. The results show that the model is able to categorise patients into the seven treatment groups previously identified [35] and demonstrate that the final classification indeed meets the initial algorithm requirements and specifications. In addition, our proposed model provides a simple, understandable rule set for classification of patients.

In recent years, several studies have been carried out investigating the application of 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 [13–16]. In particular, Abd El-Rehim et al. [46] identified and characterised five breast cancer classes, with a sixth group of only four cases also identified but considered too small for further detailed assessment. Subsequently, we investigated the stability of the proposed classification across different case sets, assay methods and data analysis procedures by investigating the effects of multiple hard-clustering methods on a breast cancer dataset [21, 51]. This led to a clear definition of cancer classes, but left many patients in a mixed-classified or unclassified group.

A different approach to hard-clustering is the use of fuzzy methodologies which have become more and more important over recent years in addressing classification problems. Specifically, fuzzy rule-based systems have been utilised to produce high classification accuracy through linguistic rulesets. As a consequence, the fuzzyQSBA algorithm was developed [29] which uses continuous fuzzy quantifiers to create the ruleset.

Using the fuzzyQSBA method together with the class assignment algorithm presented in this paper, it was possible to obtain a refinement of the seven breast cancer classes presented by Green et al. in [35]. This has led to a more 'clinically acceptable' classification, with the proportion of HER2-positive patients ranging between 12.5% and 15% and the total number of unclassified patients being only 3.5% of the available cases. The distribution and percentages of breast cancer patients into the three big classes of luminal, basal and HER2 were established in a seminal study by Sorlie et al. [6] and confirmed by subsequent papers [52–54]. The method described here has produced a breast cancer classification consistent with the proportion of cancer subtypes reported in other studies. In addition, these new subclasses have significant differences in tumour characteristics and in clinical outcome, as reported in our most recent study [45].

A linguistic rules table representing the numerical ruleset was also produced (table 3), to facilitate the decision making process for any possible future patient having been diagnosed with breast cancer. By comparing it with the expert ruleset created by clinicians (figure 3), it can be seen how easily understandable and clinically interpretable our proposed model is. As a matter of fact, the only difference concerns the HER3 biomarker, which seems to be only relevant for the classification of patients in class 2 (and is *Omit* for all other classes). Further analysis is needed to check whether the incorporation of another marker in the model (Ki67/MIB1) can make HER3 redundant.

The distinction between the class membership and the class assignment algorithms in our proposed methodology is a real strength and can facilitate the medical decision making process. First, the class membership provides an indication of each patient's likelihood to present characteristics of the specified classes. This resulting table can be directly analysed by medical experts when deciding which therapy might be the most beneficial for a particular patient. If, instead, a more clear and decisive classification is required, it is sufficient to run the class assignment algorithm to obtain indication of each class population. However, one can argue that too many variables and thresholds need to be manually passed to the proposed approach. While we acknowledge this, and accept that it may be seen as a limitation, we argue that the existence of such parameters provides the future potential to adjust the parameters to reflect different clinical priorities or external conditions.

From a medical perspective, the definition of seven classes resulting from this paper has been used as a starting point for the creation of a clinically usable tool for prospective classification (called 'NPI+'), taking into account current therapeutic strategies [45].

In conclusion, we have shown how the use of fuzzy quantifiers in subsethood based algorithm may improve both classification accuracy and interpretability of derived rulesets. Clinicians can use the linguistic ruleset to quickly assess patients tumour biology and select the most appropriate treatment regimen accordingly. A thorough external validation phase is underway, in which more data from different European centres are being collected and scored to properly assess the accuracy of our methodology and to address concerns about biases and self-testing. In the meantime, validation on a newer small breast cancer cohort has given promising results. Future work will also focus on determining whether novel markers need to be incorporated in the model itself.

References

[1] UK CR. Cancer Incidence for Common Cancers - UK Statistics; 2011. http://info.cancerresearchuk.org/cancerstats/incidence/commoncancers (accessed: 27 February 2012).

[2] Clark GM. Do We Really Need Prognostic Factors for Breast Cancer? *Breast Cancer Res Treat.* 1994;**30**:117–126.

[3] Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in Primary Breast Cancer. *Breast Cancer Res Treat.* 1992;**22**:207–219.

[4] Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A Prognostic Index in Primary Breast Cancer. *Br J Cancer*. 1982;**45**:361–366.

[5] Ellis IO, Pinder SE, Lee AH, Elston CW. 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*. 1999;**16**:202–208.

[6] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications. *Proc Natl Acad Sci U S A*. 2001;**98**:10869–10874.

[7] van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Bernards R, et al. Expression Profiling Predicts Outcome in Breast Cancer. *Breast Cancer Res*. 2003;**5**:57–58.

[8] Van De Vijver MJ, He YD, Van'T Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med*. 2002;**347**:1999–2009.

[9] Xu R, Damelin S, Nadler B, Wunsch II DC. Clustering of High-Dimensional Gene Expression Data with Feature Filtering Methods and Diffusion Maps. *Artificial Intelligence in Medicine*. 2010;**48**(2-3):91–98.

[10] Zeng T, Liu J. Mixture Classification Model Based on Clinical Markers for Breast Cancer Prognosis. *Artificial Intelligence in Medicine*. 2010;**48**(2-3):129–137.

[11] Perou CM, Sorlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular Portraits of Human Breast Tumours. *Nature*. 2000;**406**:747–752.

[12] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets. *Proc Natl Acad Sci U S A*. 2003;**100**:8418–8423.

[13] Callagy G, Cattaneo E, Daigo Y, Happerfield L, Bobrow L, Pharoah P, et al. Molecular Classification of Breast Carcinomas Using Tissue Microarrays. *Diagn Mol Pathol*. 2003;**12**:27–34.

[14] Makretsov NA, Huntsman DG, Nielsen TO, Yorida E, Peacock M, Cheang MCU, et al. Hierarchical Clustering Analysis of Tissue Microarray Immunostaining Data Identifies Prognostically Significant Groups of Breast Carcinoma. *Clin Cancer Res*. 2004;**10**:6143–6151.

[15] Jacquemier J, Ginestier C, Rougemont J, Bardou VJ, Charafe-Jauffret E, Geneix J, et al. Protein Expression Profiling Identifies Subclasses of Breast Cancer and Predicts Prognosis. *Cancer Res*. 2005;**65**:767–779.

[16] Ambrogi F, Biganzoli E, Querzoli P, Ferretti S, Boracchi P, Alberti S, et al. Molecular Subtyping of Breast Cancer from Traditional Tumor Marker Profiles Using Parallel Clustering Methods. *Clinical Cancer Research*. 2006;**12**(3):781–790.

[17] Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*. 2003;**52**:91–118.

[18] Swift S, Tucker A, Vinciotti V, Martin N, Orengo C, Liu X, et al. Consensus Clustering and Functional Interpretation of Gene-Expression Data. *Genome Biology*. 2004;**5**(11):Article R94.

[19] Filkov V, Skiena S. Integrating Microarray Data by Consensus Clustering. *International Journal on Artificial Intelligence Tools*. 2004;**13**(04):863–880.

[20] Kellam P, Liu X, Martin N, Orengo C, Swift S, Tucker A. Comparing, Contrasting and Combining Clusters in Viral Gene Expression Data. In: Proceedings of 6th Workshop on Intelligent Data Analysis in Medicine and Pharmacology; 2001. p. 56–62.

[21] Soria D, Garibaldi JM, Ambrogi F, Green AR, Powe D, Rakha E, et al. A Methodology to Identify Consensus Classes from Clustering Algorithms Applied to Immunohistochemical Data from Breast Cancer Patients. *Computers in Biology and Medicine*. 2010;**40**:318–330.

[22] Pawlak Z, Grzymala-Busse J, Slowinski R, Ziarko W. Rough Sets. *Commun ACM*. 1995;**38**(11):88–95.

[23] Grzymala-Busse JW. Knowledge Acquisition under Uncertainty – a Rough Set Approach. *Journal of Intelligent and Robotic Systems*. 1988;**1**:3–16.

[24] Parthaláin NM, Jensen R, Shen Q, Zwiggelaar R. Fuzzy-Rough Approaches for Mammographic Risk Analysis. *Intelligent Data Analysis*. 2010;**14**(2):225–244.

[25] Li R, Wang Z. Mining Classification Rules Using Rough Sets and Neural Networks. *European Journal of Operational Research*. 2004;**157**(2):439–448.

[26] Soria D, Garibaldi JM, Biganzoli E, Ellis IO. A Comparison of Three Different Methods for Classification of Breast Cancer Data. In: Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on. IEEE Computer Society; 2008. p. 619–624.

[27] Yen J, Wang L. Simplifying Fuzzy Rule-Based Models Using Orthogonal Transformation Methods. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*. 1999;**29**(1):13–24.

[28] Ishibuchi H, Nakashima T. Effect of Rule Weights in Fuzzy Rule-Based Classification Systems. *Fuzzy Systems, IEEE Transactions on*. 2001;**9**(4):506–515.

[29] Rasmani KA, Shen Q. Subsethood-based Fuzzy Modelling and Classification. In: Proceeding of the 2004 UK Workshop on Computational Intelligence. Citeseer; 2004. p. 181–188.

[30] Chang PC, Liu CH. A TSK Type Fuzzy Rule Based System for Stock Price Prediction. *Expert Systems with Applications*. 2008;**34**(1):135–144.

[31] Ishibuchi H, Yamamoto T. Rule Weight Specification in Fuzzy Rule-Based Classification Systems. *Fuzzy Systems, IEEE Transactions on*. 2005;**13**(4):428–435.

[32] Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco, CA: Morgan Kaufman; 2005.

[33] Rasmani KA, Garibaldi JM, Shen Q, Ellis IO. Linguistic Rulesets Extracted from a Quantifier-Based Fuzzy Classification System. In: Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on. IEEE; 2009. p. 1204–1209.

[34] Rasmani KA, Shen Q. Modifying Weighted Fuzzy Subsethood-Based Rule Models with Fuzzy Quantifiers. In: Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on. IEEE; 2004. p. 1679–1684.

[35] Green AR, Powe DG, Rakha EA, Soria D, Lemetre C, Nolan CC, et al. Identification of Key Clinical Phenotypes of Breast Cancer Using a Minimised Panel of Protein Biomarkers; 2012. Submitted to the British Journal of Cancer.

[36] Kosko B. Fuzzy Entropy and Conditioning. *Information Sciences*. 1986;**40**(2):165–174.

[37] Bustince H, Mohedano V, Barrenechea E, Pagola M. Definition and Construction of Fuzzy DI-Subsethood Measures. *Information Sciences*. 2006;**176**(21):3190 – 3231.

[38] Chen SM, Lee SH, Lee CH. A New Method for Generating Fuzzy Rules from Numerical Data for Handling Classification Problems. *Applied Artificial Intelligence*. 2001;**15**:645–664.

[39] Rasmani KA, Shen Q. Weighted Linguistic Modelling Based on Fuzzy Subsethood Values. In: Fuzzy Systems, 2003. FUZZ'03. The 12th IEEE International Conference on. vol. 1. IEEE; 2003. p. 714–719.

[40] Quinlan JR. Induction of Decision Trees. *Machine Learning*. 1986;**1**:81–106.

[41] Yuan Y, Shaw MJ. Induction of Fuzzy Decision Trees. *Fuzzy Sets and Systems*. 1995;**69**(2):125–139.

[42] Vila MA, Cubero JC, Medina JM, Pons O. In: Using OWA Operators in Flexible Query Processing. Norwell, MA, USA: Kluwer Academic Publishers; 1997. p. 258–274.

[43] Castellano G, Fanelli A, Mencar C, Plantamura VL. Classifying Data with Interpretable Fuzzy Granulation. In: Proc Of SCIS & ISIS; 2006. p. 872–876.

[44] Garibaldi JM, Soria D, Rasmani KA. Consensus Clustering and Fuzzy Classification for Breast Cancer Prognosis. In: Proceedings 24th European Conference on Modelling and Simulation; 2010. p. 1–4.

[45] Rakha E, Soria D, Lemetre C, Green AR, Powe DG, Nolan CC, et al. Nottingham Prognostic Index Plus (NPI+): A Modern Clinical Decision Making Tool in Breast Cancer. *to appear in International Journal of Cancer*. 2012;.

[46] Abd El-Rehim DM, Ball G, Pinder SE, Rakha E, Paish C, Robertson JF, 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*. 2005;**116**:340–350.

[47] Maindonald JH, Braun WJ. *Data Analysis and Graphics Using R - An Example-Based Approach*. Cambridge University Press; 2003.

[48] Nottingham Prognostic Ltd. Company number: 0803561; 2012.

[49] Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960;**20**:37–46.

[50] Kendall MG. A New Measure of Rank Correlation. *Biometrika*. 1938;**30**(1/2):81–93.

[51] Soria D, Garibaldi JM, Ambrogi F, Lisboa PJG, Boracchi P, Biganzoli E. Clustering Breast Cancer Data by Consensus of Different Validity Indices. In: Advances in Medical, Signal and Information Processing, 2008. MEDSIP 2008. 4th IET International Conference on. IET; 2008. p. 1–4.

[52] van de Rijn M, Perou CM, Tibshirani R, Haas P, Kallioniemi O, Kononen J, et al. Expression of Cytokeratins 17 and 5 Identifies a Group of Breast Carcinomas with Poor Clinical Outcome. *American Journal of Pathology*. 2002;**161**(6):1991–1996.

[53] Abd El-Rehim DM, Pinder SE, Paish CE, Bell JA, Rampaul RS, Blamey RW, et al. Expression and Co-Expression of the Members of the Epidermal Growth Factor Receptor (EGFR) Family in Invasive Breast Carcinoma. *Br J Cancer*. 2004;**91**(8):1532–1542.

[54] Roepman P, Horlings HM, Krijgsman O, Kok M, Bueno-de Mesquita JM, Bender R, et al. Microarray-Based Determination of Estrogen Receptor, Progesterone Receptor, and HER2 Receptor Status in Breast Cancer. *Clinical Cancer Research*. 2009;**15**(22):7003–7011.

List of Figures

1 Framework of fuzzyQSBA [29] 19

2 Example of possible membership functions. Red colour used for term 'high', blue for 'low' 20

3 Expert ruleset in decision tree format 21

4 New ruleset in decision tree format 22

5 Comparison of boxplots of markers in the seven classes for the original 1073 patients and the new 238 patients data sets. 23

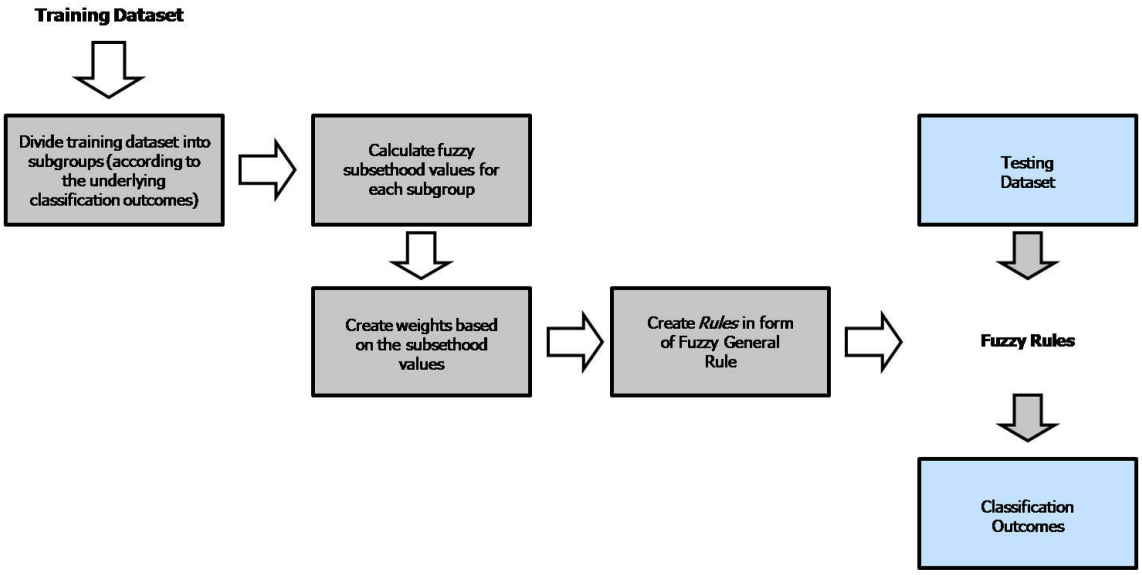


Figure 1: Framework of fuzzyQSBA [29]

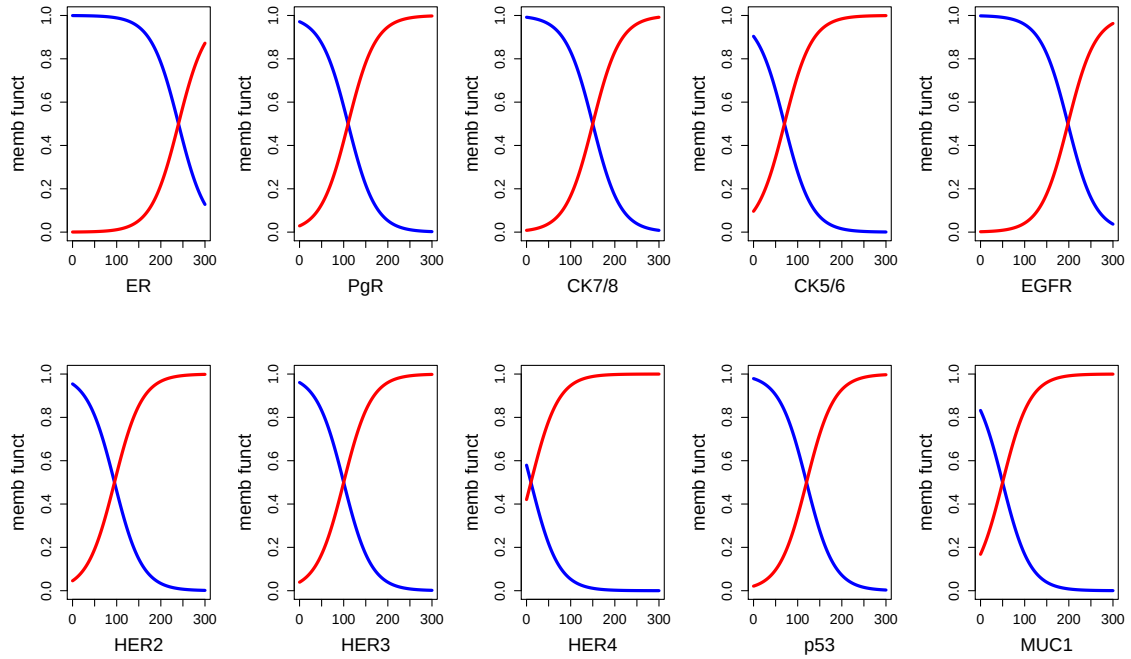


Figure 2: Example of possible membership functions. Red colour used for term 'high', blue for 'low'

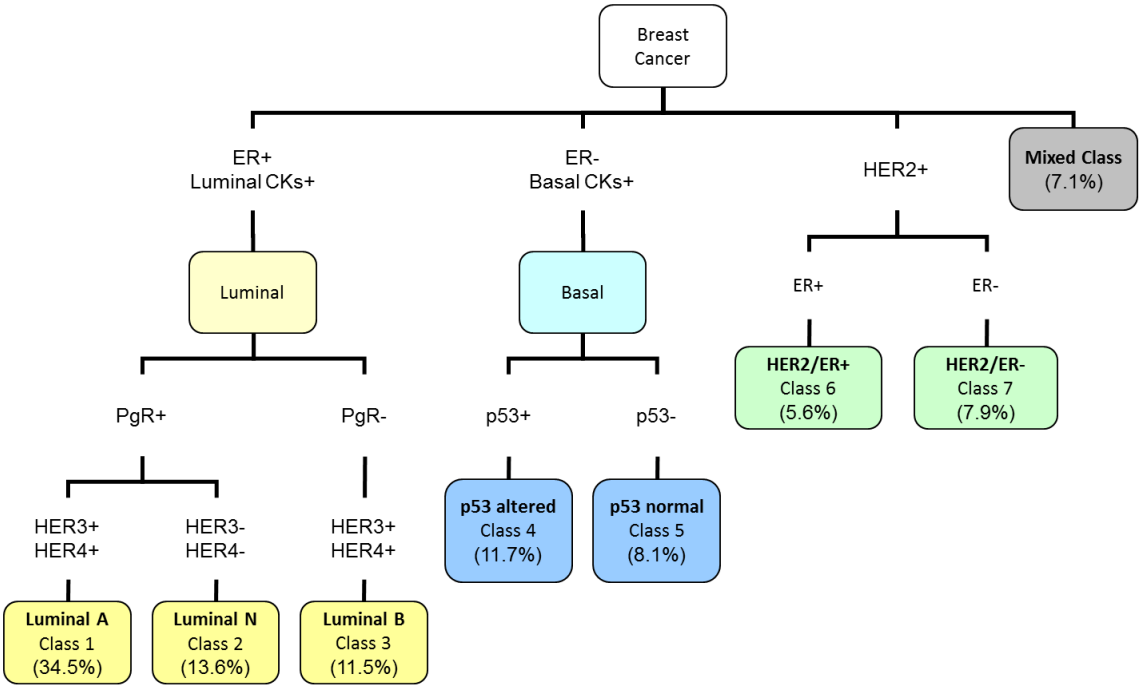


Figure 3: Expert ruleset in decision tree format

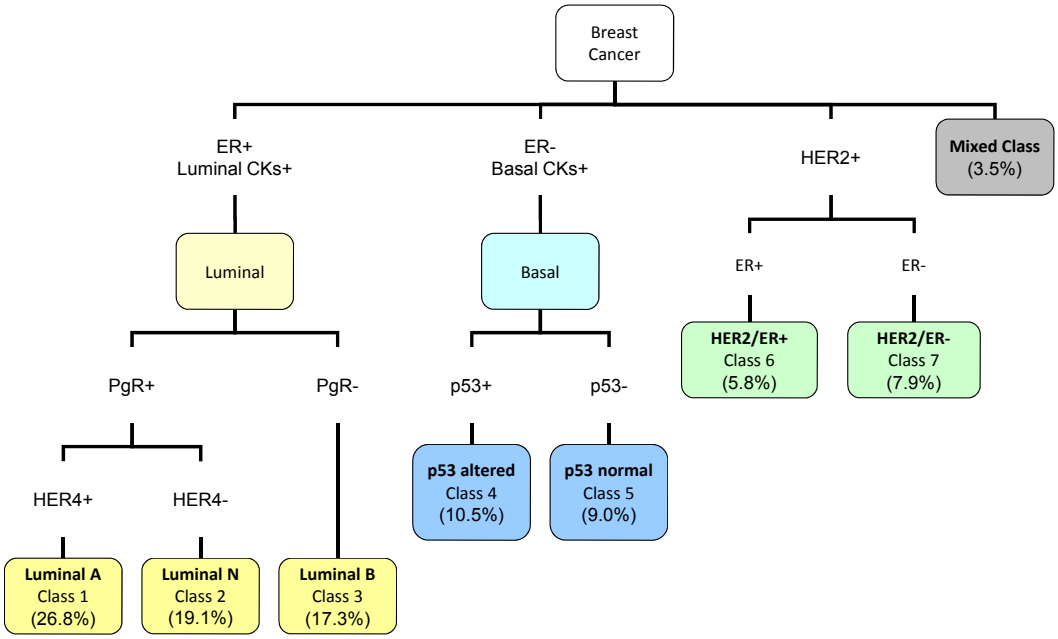
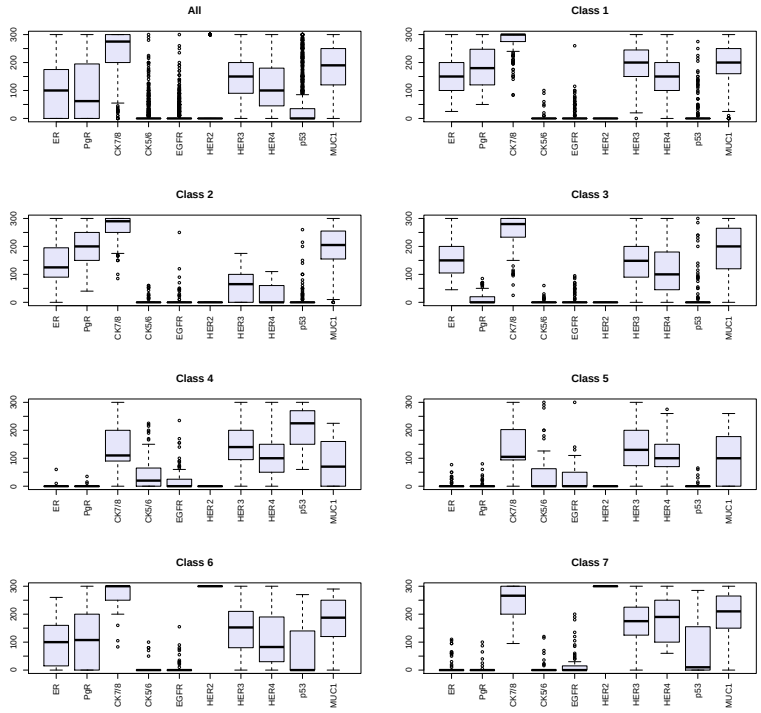
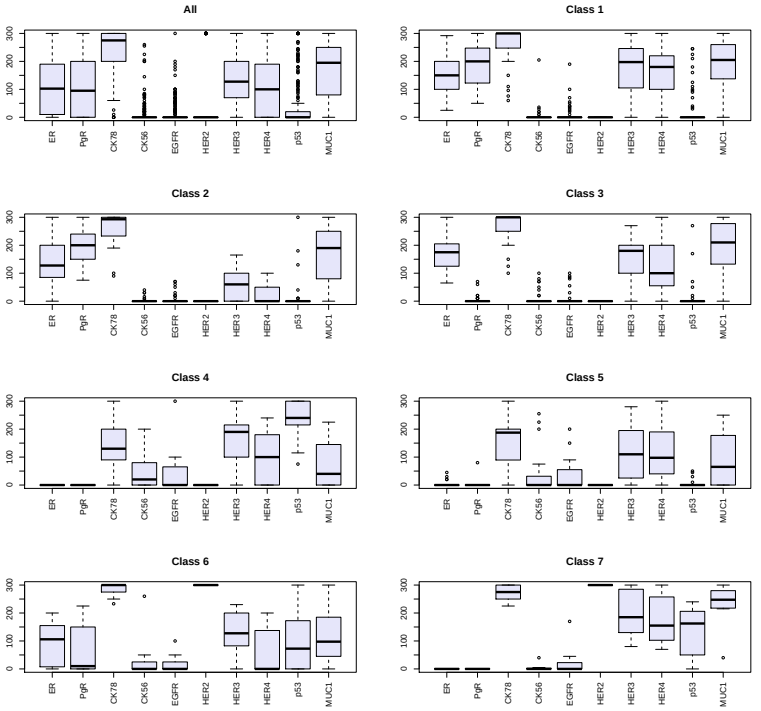


Figure 4: New ruleset in decision tree format



(a) 1073 patients



(b) 238 patients

Figure 5: Comparison of boxplots of markers in the seven classes for the original 1073 patients and the new 238 patients data sets.

List of Tables

1 Original and new class distributions ('NC' means not classified). 25

2 An example of a possible output of the algorithm. Note that patient 2879 was originally assigned to class 4, although the basal marker CK5/6 has a value of zero. This explains why the possibility value for class 4 is not particularly high. 26

3 Linguistic rules table 27

Class	1	2	3	4	5	6	7	NC
Original classification	370	146	123	126	87	60	85	76
New Classification	288	205	186	113	96	62	85	38

Table 1: Original and new class distributions ('NC' means not classified).

					Original	New Class						
Patient	ER	PgR	...	MUC1	Class	1	2	3	4	5	6	7
1625	280	0	...	235	3	0.20	0.15	0.90	0.10	0.10	0.55	0.20
2879	0	0	...	130	4	0.10	0.15	0.20	0.65	0.20	0.10	0.10
3932	200	295	...	200	1	0.85	0.10	0.20	0.15	0.20	0.10	0.05

Table 2: An example of a possible output of the algorithm. Note that patient 2879 was originally assigned to class 4, although the basal marker CK5/6 has a value of zero. This explains why the possibility value for class 4 is not particularly high.

	ER	*PgR*	*CK7/8*	*CK5/6*	*EGFR*	*HER2*	*HER3*	*HER4*	*p53*	*MUC1*
1	High	High	Omit	Omit	High	Low	Omit	High	Low	Omit
2	High	High	Omit	Omit	Omit	Low	Low	Low	Low	Omit
3	High	Low	Omit	Omit	Omit	Low	Omit	Omit	Low	Omit
4	Low	Low	Low	High	High	Low	Omit	Omit	High	Low
5	Low	Low	Low	High	High	Low	Omit	Omit	Low	Low
6	High	High	Omit	Omit	High	High	Omit	Omit	Low	Omit
7	Low	Low	Omit	High	High	High	Omit	High	High	Omit

Table 3: Linguistic rules table