Noname manuscript No.
(will be inserted by the editor)

Classification of breast cancer immunohistochemical data using semi-supervised Fuzzy c-Means

Daphne Teck Ching Lai · Jonathan M. Garibaldi · Daniele Soria · Christopher M. Roadknight

Received: date / Accepted: date

Abstract Previously, a semi-manual method was used to identify six novel and clinically useful classes in the Nottingham Tenovus Breast Cancer dataset. 663 out of 1076 patients were classified. The objectives of our work is three folds. Firstly, our primary objective is to use one single automatic method to reproduce the six classes for the 663 patients and to classify the remaining 413 patients. Secondly, we explore using semi-supervised fuzzy c-means with various distance metrics and initialisation techniques to achieve this. Thirdly, the clinical characteristics of the 413 patients are examined by comparing with the 663 patients. Our experiments use various amount of labelled data and 10-fold cross validation to reproduce and evaluate the classification. ssFCM with Euclidean distance and initialisation technique by Katsavounidis et al. [9] produced the best results. It is then used to classify the 413 patients. Visual evaluation of the 413 patients' classifications revealed common characteristics as those previously reported. Examination of clinical characteristics indicates significant associations between classification and clinical parameters. More importantly, association between classification and survival based on the survival curves is shown.

Keywords Breast Cancer · Fuzzy Clustering · Molecular Classification

Daphne Teck Ching Lai
School of Computer Science
University of Nottingham
Tel.: +44 115 95 14229
E-mail: psxdtl@nottingham.ac.uk

Jonathan M. Garibaldi
E-mail: jmg@cs.nott.ac.uk

1 Introduction

Cluster analysis of biomedical data is becoming popular and increasingly important in the prediction of illnesses or diseases that can support decision making in diagnosis and treatment. Eisen and colleagues [5] demonstrated that gene expression data can be organised into functional categories using hierarchical clustering and visual inspection of the dendrogram. This has motivated the application of such techniques on breast cancer gene expression data [14], where four breast cancer groups (ER+/luminal-like, basal-like, HER2+ and normal breast) have been identified. In a following study [17], six groups were identified where the ER+/luminal-like group was divided into three subgroups: luminal-A, B, and C. Luminal-C was later dropped [18]. Researchers moved on to cluster immunohistochemical data using hierarchical clustering, where three groups were identified in [12] and six groups in [1]. However, there has been no further investigations to address the stability of the proposed groups, that is, reproducibility of these groups, using different breast cancer datasets [18] or learning algorithms.

The Nottingham Tenovus Breast Cancer (NTBC) dataset, consisting of immunohistochemical data on 25 protein biomarkers for 1076 patients, have been clustered using hierarchical clustering into five groups, with the sixth group containing only four patients [1]. To address the stability of the proposed groups in the NTBC dataset, Soria et al. [16] used a range of techniques to reach consensus with solutions obtained from several different clustering algorithms. Using a set of manually-generated rules and clinicians' knowledge and experience, the classes are derived from those clustering solutions. The study has identified six novel and clinically useful classes of breast cancer and determined the key biomarkers that characterise these classes. As the methodology used in this study is semi-manual, it is evident that an automatic technique for the classification of breast cancer types is needed, particularly for future classification (prediction) of new patients. Out of the 1076 patients, only 663 patients were classified into the six classes. The remaining 413 patients are not classified because they were found to belong to mixed clusters based on the different clustering solutions. The classification for the remaining 413 patients will be useful as this could not only further verify the six classes and their key biomarkers but provide new insights previously not identified.

We use semi-supervised fuzzy c-means as an automatic technique to classify the NTBC dataset. Though a clustering technique, ssFCM has been demonstrated to perform classification tasks successfully using class labels [13,19, 10] with the number of clusters equals the number of classes and this is a priori. We explore using different distance metrics and initialisation techniques to achieve good classification results. We shall show our experimental results from using ssFCM with three different distance metrics Euclidean, fuzzy-weighted Mahalanobis and Mahalanobis - and with three different initialisation techniques; Simple Cluster Seeking [21], Cluster estimation [3] and technique by Katsavounidis et al. [9]. The approach which best reproduce the classification by Soria et al. is then used to classify the remaining 413 patients. The clinical characteristics of the 413 patients based on their classifications are then compared with the 663 patients for confirmation of common trends already identified and for further insights.

In this work, our objectives are of three folds:

1. To reproduce the same six classes using data of the 663 classified patients and classify the remaining 413 patients using one single method, the semi-supervised fuzzy c-means (ssFCM).
2. To explore the application of semi-supervised Fuzzy c-means in a real world problem with investigation in different initialisation techniques and distance metrics.
3. To examine the clinical characteristics of the 413 patients, comparing with those of the 663.

The paper is outlined as follows: The NTBC dataset and the methods used in this study is outline in Sect. 2. This is followed by experiments and their set-ups in Sect. 3. In Sect. 4 and 5, we present our results and discussed them respectively. We end our paper with a conclusion in Sect. 6.

2 Materials and Methods

2.1 The Nottingham Tenovus Breast Cancer Dataset

The NTBC dataset contains immunohistochemical data of 1076 patients with primary operable (stages I, II and III) invasive breast cancer between 1986 and 1998. The data is in the form of modified histochemical score (H-score) based on immunohistochemical reactivity of 25 proteins, determined using microscopical analysis. The H-score is calculated based on a semiquantitative assessment of both intensity of staining and percentage of positive cells at each intensity. The intensity of staining is quantified as score 0 to 3 correspond to negative, weak, moderate and strong positivity. The H-score ranges between 0 and 300, based on the formula below:

$$\begin{aligned}\text{H-score} &= (1 \times \%\text{ of cells with intensity } 1)\\ &+ (2 \times \%\text{ of cells with intensity } 2)\\ &+ (3 \times \%\text{ of cells with intensity } 3) \qquad (1)\end{aligned}$$

The 25 protein biomarkers used in this study are the same ones listed in [1, 16] and are shown in Table 1. The dataset also contains tumour information such as histologic grade, histologic tumour type, vascular invasion, tumour size and lymph node stage and patient information such as age and menopausal status. Survival in months from the date of primary treatment to the time of death is also recorded where patients are followed up at 3-months intervals initially, then every 6 months, then annually for a range of 1-192 months, with a median period of 58 months. The Nottingham Prognostic Index (NPI) score is also provided in the dataset and is calculated according to the formula: NPI Score = (0.2 × size) + grade + stage.

Table 1 Protein biomarkers and their dilutions.

Antibody, clone	Short name	Dilution
Luminal phenotype		
CK 7/8 [clone CAM 5.2]	CK7/8	1:2
CK 18 [clone DC 10]	CK18	1:50
CK 19 [clone BCK 108]	CK19	1:100
Basal phenotype		
CK 5/6 [clone D5/16134]	CK5/6	1:100
CK 14 [clone LL002]	CK14	1:100
SMA [clone 1A4]	Actin	1:2000
p63 ab-1 [clone4A4]	p63	1:200
Hormone receptors		
ER [clone 1D5]	ER	1:80
PgR [clone PgR 636]	PgR	1:100
AR [clone F39.4.1]	AR	1:30
EGFR family members		
EGFR [clone EGFR.113]	EGFR	1:10
c-erbB-2	HER2	1:250
c-erbB-3 [clone RTJ1]	HER3	1:20
c-erbB-4 [clone HFR1]	HER4	6:4
Tumour suppressor genes		
p53 [clone DO7]	p53	1:50
nBRCA1 Ab-1 [clone MS110]	nBRCA1	1:150
Anti-FHIT [clone ZR44]	FHIT	1:600
Cell adhesion molecules		
Anti E-cad [clone HECD-1]	E-cad	1:10/20
Anti P-cad [clone 56]	P-cad	1:200
Mucins		
NCL-Muc-1 [clone Ma695]	MUC1	1:300
NCL-Muc-1 core [clone Ma552]	MUC1co	1:250
NCL muc2 [clone Ccp58]	MUC2	1:250
Apocrine differentiation		
Anti-GCDFP-15	GCDFP	1:30
Neuroendocrine differentiation		
Chromogranin A [clone DAK-A3]	Chromo	1:100
Synaptophysin [clone SY38]	Synapto	1:30

Table 2 Number of data patterns in each class and the number of not classified (n.c) and classified (c) data patterns according to classification by Soria et al.

class 1	class 2	class 3	class 4	class 5	class 6	n.c	c
202	153	80	82	69	77	413	663

According to classification by Soria et al. [16] (which we shall refer to Soria's classification in short), 663 data patterns are classified while 413 remains not classified, as shown in Table 2.

2.2 Semi-supervised Fuzzy c-means

Fuzzy c-means is a clustering algorithm which allows a data pattern to belong to more than one cluster, giving a more realistic representation of data than a binary approach. This is particularly useful in clustering biomedical data as there are often no clear boundaries separating the classes. Membership values indicate the degree of belongingness a data pattern has to clusters and thus, determine the cluster a data pattern is assigned to. For each data pattern, membership values to each cluster range between 0 and 1 and the sum of membership values for all clusters must equal to one. A high membership value to a cluster means high possibility of belonging to this cluster.

Semi-supervised Fuzzy c-Means (ssFCM) use some labelled data patterns in the dataset to guide the identification of similar data patterns. This can be very valuable when some cases can be labelled. Labelled data patterns are often sparse and they are time-consuming and labour-intensive to collect. Also, human errors can be introduced when labels are given manually.

The ability of ssFCM to represent data in more than one clusters using membership values and to learn from labelled data patterns, which we can obtained from Soria's classification deemed it a suitable technique to use for this work. Also, ssFCM has been successfully applied in areas of biomedicine such as in [20]. Another benefit of using ssFCM is that clustering is not a statistical inference technique and is not affected by the assumptions of normal distribution [7], which is suitable on the NTBC dataset with features that are non-normally distributed. In this work, we adopt this practice with the number of clusters c equals to the six classes identified by Soria et al. [16].

2.2.1 Algorithm

Pedrycz and Waletzky [13] introduced the following ssFCM objective function containing unsupervised learning in the first term and supervised learning in the second term:

$$J = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{p} d_{ik}^{2} + \alpha \sum_{i=1}^{c}\sum_{k=1}^{N} (u_{ik} - f_{ik} b_k)^p d_{ik}^{2}, \tag{2}$$

where u_{ik} is the membership value of data pattern k in cluster i, d_{ik} the distance between data pattern k and cluster centre v_i, f_{ik} the membership value of labelled data pattern k in cluster i, b_k indicates if data pattern k is labelled, c is the number of clusters, N the number of data patterns in the dataset and p is the fuzzifier parameter (which is commonly 2) and α is a scaling parameter for maintaining balance between the supervised and unsupervised learning components. The authors recommended α to be proportional to N/M where M is the number of labelled data.

The algorithm iteratively calculates the cluster centres and the membership matrix U containing u_{ik} to minimise the objective function until a termination criterion is satisfied. In this work, we use ssFCM by Pedrycz and Waletzky

[13] because it has been shown to produce good classification results. The algorithm is summarised as follows:

1. Initialise labelled data membership matrix $\mathbf{F}$ and initial membership matrix $\mathbf{U^0}$
2. Calculate cluster centres $\mathbf{V} = [\mathbf{v_i}]$ with $\mathbf{U}$ using equation:
$$\mathbf{v_i} = \frac{\sum_{k=1}^{N} u_{ik}^2 \mathbf{x_k}}{\sum_{k=1}^{N} u_{ik}^2} \tag{3}$$
3. Update partition matrix, $\mathbf{U}$ using equation :
$$u_{ij} = \frac{1}{1+\alpha} \left\{ \frac{1 + \alpha(1 - b_j \sum_{l=1}^{c} f_{lj})}{\sum_{l=1}^{c} (\frac{d_{ij}}{d_{lj}})^2} + \alpha f_{ij} b_j \right\} \tag{4}$$
4. If $||\mathbf{U'} - \mathbf{U}|| < \epsilon$, stop. Else, go to step 2 with $\mathbf{U} = \mathbf{U'}$

2.2.2 Distance Metrics

Distance metrics are important in Fuzzy c-means as they are used to measure similarity. The degree of similarity enables us to determine how strongly a data pattern belong to a certain group. The better a distance metric is in representing the structure of the data, the more accurate is its measure of similarity.

Mahalanobis The Mahalanobis distance is formally defined [11] as:

$$d_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)} \tag{5}$$

It is the distance between a vector $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_n})^\mathbf{T}$ which belong to a group of vectors with a mean $\mu = (\mu_\mathbf{1}, \mu_\mathbf{2}, ...\mu_\mathbf{n})^\mathbf{T}$ and $\mathbf{S}$ is the covariance matrix of the group. The inverse covariance matrix, $\mathbf{S^{-1}}$ normalises dimensions of different scales, preventing dominance from dimensions with greater scales. Thus, it is scale-invariant. It forms hyperellipsoidal clusters.

The Mahalanobis distance in ssFCM [13] is a fuzzy-weighted form of Mahalanobis distance (which we shall address as fuzzy Mahalanobis in short) as it takes into account the membership in the calculation of the covariance matrices. The introduction of fuzzy weights u_{ik} in the covariance matrix was proposed by Gustafson and Kessel [6] to adapt the distance metric to the shape of clusters. The fuzzy Mahalanobis distance is computed as follows:

$$d_M^2(i,k) = (\mathbf{x_k} - \mathbf{v_i})^T \mathbf{M_i} (\mathbf{x_k} - \mathbf{v_i}) \tag{6}$$

where M_i is a positive definite matrix, its inverse defined as:

$$\mathbf{M_i}^{-1} = \left[\frac{1}{\rho_i det(\mathbf{P_i})} \right]^{\frac{1}{n}} \mathbf{P_i} \tag{7}$$

and $\mathbf{P_i}$ is the fuzzy covariance matrices defined as:

$$\mathbf{P_i} = \frac{\sum_{k=1}^{N} u_{ik}^2 (\mathbf{x_k} - \mathbf{v_i})(\mathbf{x_k} - \mathbf{v_i})^T}{\sum_{k=1}^{N} u_{ik}^2} \tag{8}$$

Euclidean The Euclidean distance metric forms spherical clusters and does not reflect scale differences among dimensions in high-dimensional datasets. It is computed as follows:

$$d_E^2(i,k) = ||\mathbf{x_k} - \mathbf{v_i}||^2$$

2.3 Initialisation Techniques

Initialisation techniques uses information from the data to give a more guided initialisation than random initialisation. Also, memberships from available labelled data, particularly when availability is low, may not provide a good initialisation. In [10], Li et al. applied Fuzzy c-means for initialisation before performing classification with ssFCM. In this work, we use Simple Cluster Seeking (SCS), initialisation technique by Katsavoundis et al. [9] (KKZ) and Cluster Estimation (CE) prior to the classification task.

2.3.1 Simple cluster seeking initialisation (SCS)

The SCS technique [21] is summarised as follows (as described in [8]:

1. The first pattern is initialised as the first cluster centre, i.e $\mathbf{v_1} = \mathbf{x_1}$.
2. For $k = 2, ..., N$, $\mathbf{x_k}$ is the next cluster centre if $||\mathbf{x_k} - \mathbf{v_i}|| > \rho$ for all existing cluster centres, where ρ is a threshold. When c cluster centres are initialised, stop. Else, decrease the value of ρ and repeat the steps.

2.3.2 Katsavounidis et al. initialisation (KKZ)

The KKZ technique [9] takes on the following steps as described in [2]:

1. Initialise the first cluster centre with the data pattern that has the maximum norm, $\mathbf{v_1} = argmax||\mathbf{x_k}||$.
2. Initialise the second cluster centre with the data pattern furthest from $\mathbf{v_1}$.
3. Compute the minimum distances between the remaining points with all initialised cluster centres. The data pattern with the largest value of these minimum distances are chosen as the next cluster centre.
4. Repeat step 3 until all cluster centres are found.

2.3.3 Cluster estimation technique (CE)

The cluster estimation technique [3] estimates both the number and location of cluster centres by specifying its neighbourhood size. Based on the number of neighbouring patterns, a potential value is calculated as follows:

$$P_i = \sum_{j=1}^{n} e^{-\alpha||\mathbf{x_i} - \mathbf{x_j}||^2} \tag{9}$$

Table 3 Classification results of ssFCM using Euclidean (E), Mahalanobis(M) and fuzzy Mahalanobis (FM) distances.

Dist.	10%	20%	30%	40%	50%	60%
E	0.965±0.013	0.977±0.007	0.984±0.005	0.987±0.004	0.991±0.004	0.992±0.003
M	0.775±0.025	0.859±0.017	0.897±0.012	0.925±0.010	0.942±0.009	0.957±0.007
FM	0.453±0.037	0.537±0.035	0.606±0.033	0.691±0.040	0.757±0.034	0.828±0.032

where $\alpha = \frac{4}{r_a^2}$. The pattern with the highest potential value becomes the first cluster centre. Eq. (9) is then revised to calculate the potential of patterns to be centres of other clusters as shown below:

$$P_i \Leftarrow P_i - P_k^* e^{-\beta ||\mathbf{x_i} - \mathbf{x_k^*}||^2} \qquad (10)$$

where $\beta = \frac{4}{r_b^2}$, $\mathbf{x_k^*}$ is the latest obtained cluster centre and P_k^* its potential. The positive constants r_a and r_b are radius defining their respective neighbourhoods. The author recommended that r_b=1.25r_a.

3 Experiments and set-up

3.1 Classification using random stratified sampling of labelled data

We classify NTBC using ssFCM with Euclidean, Mahalanobis and fuzzy Mahalanobis distances, with the purpose to explore how well ssFCM can find the substructures as those with Soria's [16]. Varying amounts of labelled data are experimented with; 10%, 20%, 30%, 40%, 50% and 60% of the 663 classified data patterns. To select data patterns to be labelled, random stratified sampling is applied across the six classes. We experiment with each varying amount across 100 different sets of labelled data.

To determine the class of a data pattern $\mathbf{x_k}$, we choose the class with the highest membership value. To evaluate the accuracy of the algorithm, the classes assigned by ssFCM to the 663 data patterns are then compared with Soria's classification [16] and the matches counted and divided by 663. An average is then taken across the 100 runs.

3.2 Classification using cross validation

We train and test the algorithm using 10-fold cross validation where 90% of the 663 data patterns are training data and the remaining 10% is testing data. The algorithm is run 30 times on randomly selected labelled data, across varying amount of labelled data; 10%, 20%, 30%, 40%, 50% and 60% of training data using three distance metrics, Euclidean, Mahalanobis and fuzzy Mahalanobis. The classification result obtained from the training process is then used to initialise the algorithm for the testing process. For evaluation, only matches of testing data are counted and divided by the number of testing data and an average is taken across the 30 runs for all 10 folds. This average indicates the agreement level of our solutions with Soria's classification [16].

The reason two evaluation settings have been used is because in many ssFCM literatures, evaluation is performed as shown in Sect. 3.1 but this is considered to produce optimistic results as it has not been tested on unseen data. Hence, the need for using the cross validation technique. For completeness, we demonstrate both evaluation settings.

3.2.1 Initialising membership

Soria's classification [16] is used to generate membership values which are then used to initialise the supervision matrix $\mathbf{F}$ which contains membership values for labelled data. Instead of using random initialisation, we use the supervision matrix $\mathbf{F}$ to initialise the membership matrix $\mathbf{U^0}$. This should give a better start rather than a random one. Only data patterns classified by Soria et al. [16] are used and the 413 data patterns not classified are disregarded as we do not have labels for them. To initialise membership values in $\mathbf{F}$, the selected labelled data patterns belonging to their respective classes will be given a membership of 0.9 and (1-0.9)/(6-1)=0.02 for classes they do not belong to. The high 0.9 membership value is arbitrarily chosen to indicate a data pattern's high possibility of belonging to the class while a 0.02 value indicates otherwise. Unlabelled data patterns have a membership value of $1/c \approx 0.1667$ to indicate equal possibility of belonging to the classes.

3.2.2 Configuration of ssFCM

In the original ssFCM [13], all data patterns are assigned memberships based on their given labels and stored in $\mathbf{F}$. They are then selected to be labelled or unlabelled for the algorithm using the boolean vector $\mathbf{b}$ in (2). In our case, we have selected the labelled data for the algorithm and generated their memberships prior to running the algorithm. We set our $\mathbf{F} = \mathbf{U^0}$, where they contain memberships of both labelled and unlabelled data and b_k is 1 for all k (in (4)). The α value is set to be N/M where M is the number of labelled data.

3.2.3 Using initialisation techniques

Initialisation techniques, SCS, KKZ and CE are used to initialise initial cluster centres, $\mathbf{V^0}$, instead of using (3). Similar initialisation procedures as described above are carried out to initialise memberships of supervision matrix, $\mathbf{F}$.

3.3 Breast cancer type classification for the 413 patients

We train using the 663 labelled data patterns and ssFCM was able to retain the whole labelled data completely, meaning 100% training accuracy. As all the data patterns are labelled in this training process, it is actually a completely supervised process. We classify the 413 unlabelled data patterns based on the model built from the training process. The class labels assigned to these data patterns are based on the highest membership value it has to a class. These assigned class labels are refer as the classification of the 413 data patterns. As we do not have prior labels to evaluate the correctness of the classification, we evaluate by visually comparing boxplots and biplots of biomarker distributions across the 6 classes with those by Soria and colleagues [16] and perform clinical evaluation by investigating in the correlation between the class distributions and clinical parameters.

Table 4 Classification results of ssFCM using Euclidean (E), Mahalanobis(M) and fuzzy Mahalanobis (FM) distances based on cross validation.

Dist.	10%	20%	30%	40%	50%	60%
E	0.961±0.020	0.969±0.019	0.972±0.018	0.975±0.016	0.976±0.015	0.978±0.015
M	0.750±0.057	0.814±0.056	0.846±0.051	0.860±0.051	0.869±0.049	0.876±0.049
FM	0.351±0.074	0.389±0.065	0.419±0.053	0.437±0.049	0.461±0.048	0.480±0.055

Table 5 Classification results of ssFCM using Euclidean (E)and Mahalanobis (M) distances and initialisation techniques SCS, KKZ and CE.

Dist.	10%	20%	30%	40%	50%	60%
E-SCS	*0.964±0.020*	0.969±0.019	0.972±0.018	*0.976±0.016*	0.976±0.015	0.978±0.015
E-KKZ	*0.965±0.019*	*0.971±0.019*	*0.974±0.017*	*0.976±0.016*	*0.977±0.015*	*0.979±0.015*
E-CE	0.959±0.020	0.967±0.019	0.970±0.018	0.974±0.016	0.975±0.016	0.978±0.016
M-SCS	0.749±0.056	0.814±0.055	0.846±0.050	*0.861±0.051*	*0.870±0.049*	*0.878±0.049*
M-KKZ	*0.751±0.055*	0.814±0.055	0.846±0.050	*0.861±0.050*	*0.870±0.048*	*0.878±0.049*
M-CE	0.749±0.056	0.813±0.054	0.846±0.050	0.860±0.051	*0.870±0.049*	*0.878±0.049*

4 Results

4.1 Classification using random stratified sampling

Table 3 shows the classification results using ssFCM with Euclidean, Mahalanobis and fuzzy Mahalanobis distances based on the average rate of matching class assignments with Soria's classification followed by ± *standard deviation*. ssFCM with Euclidean distance gave the best result, achieving 0.97 agreement with 10% labelled data. With 50% labelled data or more, almost complete agreement was achieved. This is no surprise as the distance metric used in the clustering techniques by Soria et al. [16] is Euclidean distance. Interestingly, higher accuracy rates was found using Mahalanobis distance than fuzzy Mahalanobis distance. To the best of our knowledge such trend with poorer accuracy using fuzzy Mahalanobis than Mahalanobis distances have never been reported, despite it being widely used in Fuzzy c-means. In [13], ssFCM with fuzzy Mahalanobis distance has produced higher accuracy than with original Mahalanobis distance for UCI Iris dataset and XOR dataset. However, in a separate unpublished study with five UCI datasets (Ionosphere, Page Blocks, Pima Indian Diabetes (PID), Wine and Wisconsin Original Breast Cancer (WOBC)), we found that the fuzzy Mahalanobis distance perform less

Table 6 Confusion matrix showing number of agreement and disagreement between ssFCM-E-KKZ (row) and ssFCM-M-KKZ (column) and their class distributions.

	class1	class2	class3	class4	class5	class6	total
class1	78	5	3	1	1	3	91
class2	21	70	8	4	5	5	113
class3	11	1	47	2	4	1	66
class4	0	0	2	10	0	3	15
class5	2	3	7	0	45	1	58
class6	12	4	2	5	3	44	70
total	124	83	69	22	58	57	413

favourably than Mahalanobis distance for PID, Wine and WOBC datasets, suggesting that fuzzy Mahalanobis distance do not always produce better results than the original Mahalanobis distance for all datasets.

4.2 Classification using cross validation

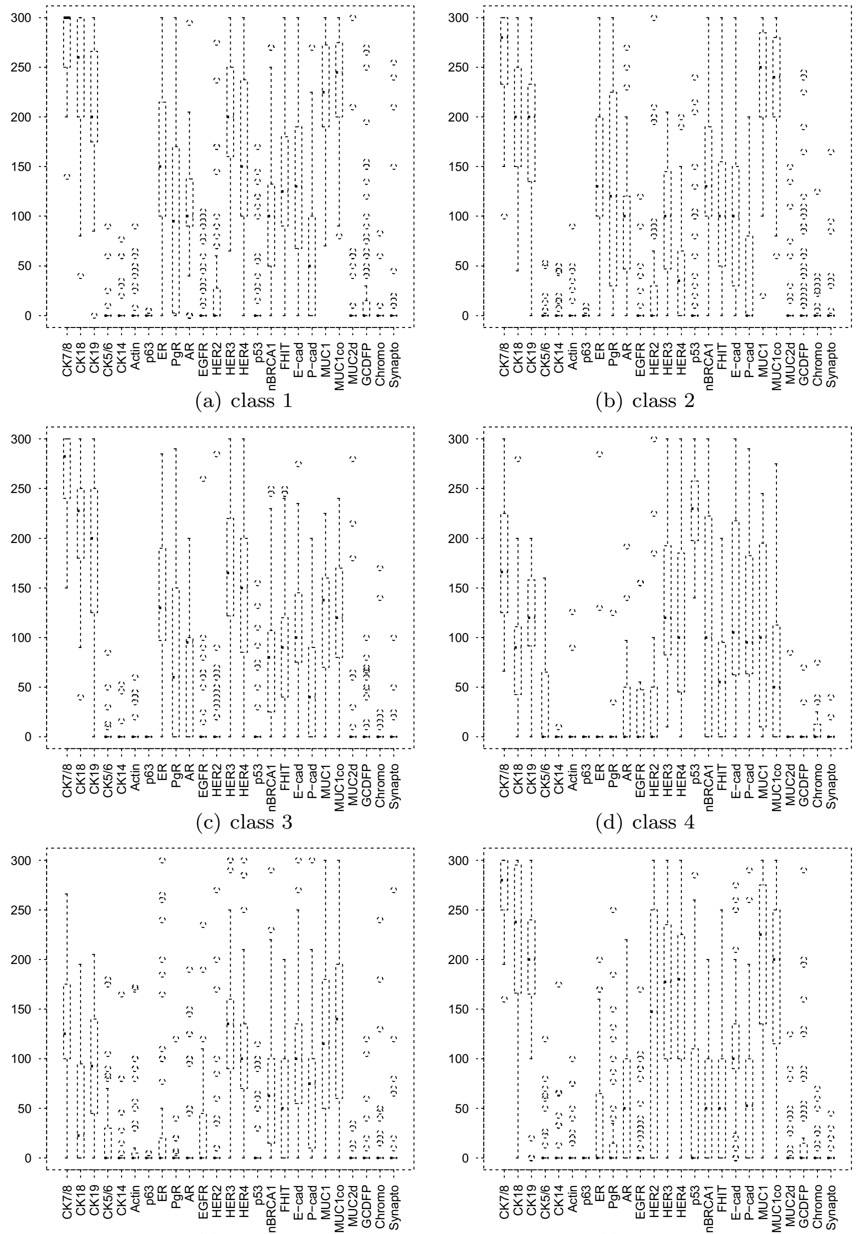
The classification result of ssFCM using Euclidean, Mahalanobis and fuzzy Mahalanobis distances in Table 4 shows high agreement with Soria's classification using ssFCM with Euclidean distance when classifying breast cancer types of the 413 patients. ssFCM with Mahalanobis distance performed moderately well but with fuzzy Mahalanobis distance, agreement level was less than half even with 60% labelled data.

Table 5 shows the classification result of ssFCM using Euclidean and Mahalanobis distance metrics with initialisation techniques; SCS, KKZ and CE. The results improved slightly using initialisation techniques (indicated in italics), particularly with KKZ for ssFCM with Euclidean distance. Slight improvement was also found in ssFCM with Mahalanobis distance with SCS and KKZ. While the increase in agreement is small, they are crucial in achieving more accurate classification to support medical decision making. The results of ssFCM with fuzzy Mahalanobis are poor from Table 4, the results are not displayed and its further investation is stopped. Instead, we shall focus on ssFCM with Euclidean distance and KKZ and with Mahalanobis distance and KKZ, which we will refer as ssFCM-E-KKZ and ssFCM-M-KKZ respectively.

Also, ssFCM with Euclidean distance and ssFCM-E-KKZ produced better classification results of NTBC (with agreement of 0.978 and 0.979 respectively using 60% labelled data) than using classifiers C4.5, Multilayer Perceptron and Naive Bayes (with agreement of 0.878, 0.976 and 0.869 respectively [15]). We continue using ssFCM-M-KKZ for further investigation to compare with ssFCM-E-KKZ.

4.3 Classification for 413 patients.

A confusion matrix is drawn up to show the number of agreement and disagreement between the classifications of the two ssFCMs in Table 6. The table also shows the class distributions of the classifications. There appears to be higher number of disagreement with class 1 of ssFCM-M-KKZ with class 2, 3 and 6 of ssFCM-E-KKZ. Using Cohen's Kappa and weighted Kappa Index



(a) class 1

(b) class 2

(c) class 3

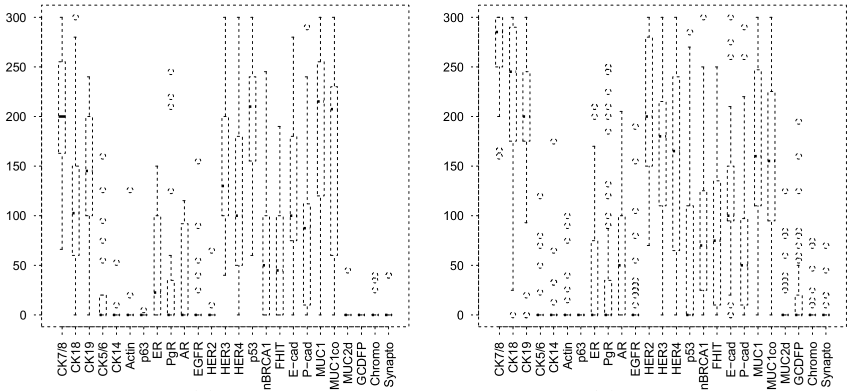
(d) class 4

(e) class 5

(f) class 6

Fig. 1 Boxplots showing statistical summaries of all biomarkers for the six classes obtained from ssFCM-E-KKZ.

as measures of agreement between the two classifications, values of 0.643 and 0.668 were obtained respectively.



(a) class 4

(b) class 6

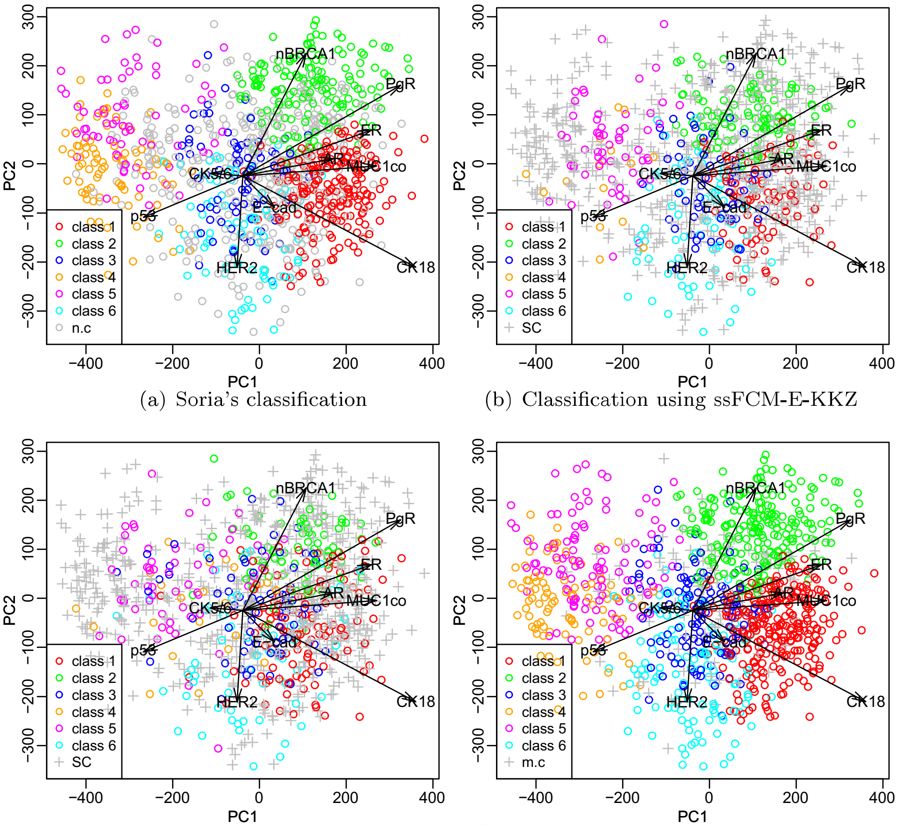
Fig. 2 Boxplots showing statistical summaries of all biomarkers for a) class 4 and b) class 6 obtained from ssFCM-M-KKZ.

4.3.1 Visual inspection

We use visual inspection of boxplots and biplots in Figures 1 and 3 to compare how closely distribution of key biomarkers agree with Soria's classification. The boxplots of ssFCM-E-KKZ and ssFCM-M-KKZ are quite similar except for boxplots of ER in classes 4 and 5 and of HER2 in class 6. According to Soria's classification, ER in classes 4 and 5 has low expression, which is reflected in ssFCM-E-KKZ's classification (see Figures 1(d) and 1(e))) but in ssFCM-M-KKZ, classes 4 and 5 have a high upper quartile range of 150 even though their medians are 25 and zero respectively (see boxplot for class 4 in Figure 2(a)). Thus, boxplots of classes 4 and 5 from ssFCM-E-KKZ more closely reflect the characteristics of triple negative breast cancer.

Comparing the boxplots of Soria's classification with ssFCM-E-KKZ and ssFCM-M-KKZ, the boxplots of the two ssFCMs show a similar general trend to Soria's classification. But, the degree of dispersion in the boxplots is found higher with the ssFCMs' classifications for some biomarkers such as PgR in class 1 and 2. This explains the mixture of classes the 413 patients belong to, as stated in [16] and hence, the difficulty in classifying these patients. There is little number of patients with high CK14 H-score in the population of 413 patients to distinguish class 4 and 5 from the other classes. Instead, biomarkers CK5/6 and p53 help contribute to make the distinction.

Figure 3 shows the biplots of Soria's classification, ssFCM-E-KKZ and ssFCM-M-KKZ which are constructed using the first two principal components (PC1 and PC2 on the biplots) of Principal Component Analysis (PCA). Note that no feature reduction was done using PCA but rather the first two components are used solely for visualisation. The biplot of ssFCM-E-KKZ(Figure 3(b)) have more compact clusters that have clearer separation which more closely resemble that of Soria's classification (Figure 3(a)) than that of ssFCM-M-KKZ (Figure 3(c)) where the clusters are more scattered and overlapped. Figure 3(d) shows a biplot of combination of Soria's classification and agreeing solutions of ssFCM-E-KKZ and ssFCM-M-KKZ. The patients that belong to



(a) Soria's classification

(b) Classification using ssFCM-E-KKZ

(c) Classification using ssFCM-M-KKZ

(d) Combined Soria and agreeing ssFCM classifications

Fig. 3 Biplots showing the six classes of Soria's classification (SC) [16] and not classified (n.c) patients in (a) and the remaining 413 patients using ssFCM-E-KKZ in (b) and ssFCM-M-KKZ in (c) and combined Soria's classification with agreeing solutions of ssFCM-E-KKZ and ssFCM-M-KKZ and mixed class (m.c) patients in (d).

mixed classes (denoted m.c in the figure) are shown as points at the edge/s of two or more clusters.

4.3.2 Analysis of cluster centres

The key biomarkers that characterise the six classes can be identified from analysis of the cluster centres generated by the two ssFCMs. The biomarkers with standard deviation of the six cluster centres less than a value of 30 and that do not contain very different values have been removed as they do not help to discriminate between the classes. The values that are underlined in Table 7 indicate high expression of the biomarker. We can observed very high p53 and HER2 expressions which confirms Soria's classification for classes 4 and 6 characteristics respectively. The high ER values in classes 1-3 separate them from classes 4-6. We can also observe what differentiates classes 1-3.

Table 7 Cluster centres for the six classes generated from the ssFCMs.

	CK7/8	CK18	CK19	CK5/6	CK14	ER	PgR	AR	HER2	HER3	HER4	p53	nBR1	MUC1	MUC1c
ssFCM-E-KKZ															
1	$\underline{286.8}$	$\underline{252.1}$	$\underline{224.8}$	1.0	0.8	$\underline{149.9}$	$\underline{170.0}$	$\underline{102.9}$	14.1	$\underline{197.7}$	$\underline{161.0}$	10.5	97.1	$\underline{216.2}$	$\underline{226.9}$
2	$\underline{270.7}$	$\underline{201.5}$	$\underline{208.5}$	2.3	2.7	$\underline{135.0}$	$\underline{178.6}$	$\underline{101.4}$	14.5	73.4	34.2	8.1	$\underline{166.2}$	$\underline{215.1}$	$\underline{227.2}$
3	$\underline{267.9}$	$\underline{212.6}$	$\underline{190.9}$	4.2	1.9	$\underline{147.2}$	72.7	76.7	26.5	$\underline{155.0}$	$\underline{135.4}$	18.5	84.4	81.5	84.5
4	123.0	39.8	76.9	$\underline{37.6}$	$\underline{30.2}$	6.9	3.6	7.2	7.7	$\underline{157.2}$	112.5	$\underline{225.8}$	52.6	86.1	85.8
5	113.3	40.0	76.8	$\underline{36.2}$	$\underline{24.3}$	20.8	13.5	12.7	9.7	109.2	88.0	28.1	77.4	91.9	83.2
6	$\underline{259.7}$	$\underline{210.0}$	$\underline{197.6}$	6.5	4.5	34.9	24.2	53.7	$\underline{180.1}$	$\underline{169.1}$	$\underline{159.4}$	78.0	61.8	$\underline{220.0}$	$\underline{210.7}$
sd	79.62	94.15	67.40	17.38	12.99	68.23	78.99	42.18	67.94	44.74	48.50	84.50	40.58	71.62	75.34
ssFCM-M-KKZ															
1	$\underline{274.2}$	$\underline{233.8}$	$\underline{210.7}$	3.6	2.4	$\underline{135.1}$	$\underline{152.7}$	$\underline{93.1}$	18.8	$\underline{185.6}$	$\underline{151.7}$	19.4	95.8	$\underline{204.7}$	$\underline{213.3}$
2	$\underline{261.9}$	$\underline{193.1}$	$\underline{199.9}$	5.2	4.2	$\underline{120.8}$	$\underline{155.1}$	$\underline{91.4}$	19.6	89.4	56.7	18.1	$\underline{150.2}$	$\underline{197.8}$	$\underline{209.7}$
3	$\underline{256.3}$	$\underline{199.4}$	$\underline{184.3}$	7.2	4.2	$\underline{132.5}$	77.2	77.1	30.2	$\underline{148.9}$	$\underline{123.7}$	27.5	92.4	113.0	116.2
4	181.0	109.1	125.2	$\underline{23.6}$	$\underline{18.8}$	53.6	45.5	37.6	19.0	$\underline{153.1}$	$\underline{115.8}$	$\underline{144.6}$	75.4	132.5	132.7
5	176.2	113.2	125.0	$\underline{27.0}$	$\underline{18.3}$	66.1	56.9	43.7	22.0	$\underline{128.4}$	98.9	31.4	91.0	139.7	133.0
6	$\underline{249.5}$	$\underline{194.5}$	$\underline{187.7}$	8.3	5.6	59.9	54.0	60.6	$\underline{136.1}$	$\underline{156.3}$	$\underline{137.5}$	67.9	79.0	$\underline{203.4}$	$\underline{200.6}$
sd	43.06	50.81	37.62	10.10	7.52	38.61	50.39	23.77	46.80	32.29	33.40	49.11	27.12	41.29	44.70

Class 3 has much lower PgR and AR values than classes 1 and 2, and class 2 has much lower HER3 and HER4 values. These expressions coincide with Soria et al's proposed summary of classes with class interpretations (see Fig. 5 in [16]). We observed that the cluster centres generated by ssFCM-E-KKZ have higher discriminating power than those generated by ssFCM-M-KKZ as there is greater difference between values that are underlined from those that are not in ssFCM-E-KKZ than ssFCM-M-KKZ. This is also due to higher standard deviation for each biomarker in ssFCM-E-KKZ than ssFCM-M-KKZ.

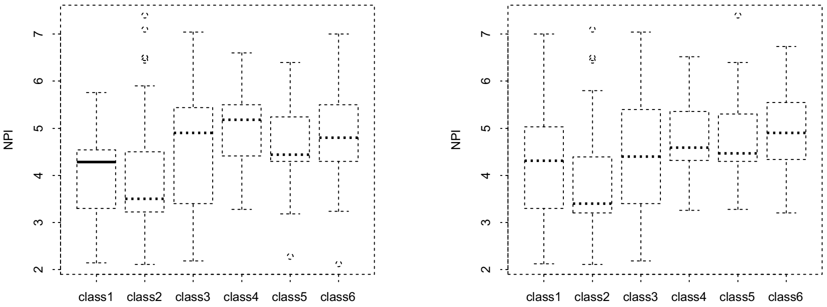
4.3.3 Clinical evaluation

Clinical evaluation is conducted by investigating the association between the classes with respect to clinical parameters such as age, tumour grade, etc. are shown in Table 8. The ϕ coefficient values from the table shows significant associations between classes. Classification from ssFCM-M-KKZ (in brackets) has slightly higher ϕ coefficient values. The boxplots of class distribution with respect to NPI values in Figure 4 show a clear distinction in NPI values between classes 1-2 and classes 4-6 but in class 3, there is a higher degree of dispersion which overlaps between the two groups using ssFCM-E-KKZ. Using ssFCM-M-KKZ, there is a clear distinction between class 2 and classes 4-6 but, overlapping occurs in class 1 and 3.

In our survival analysis in Figure 5, we have removed surviving patients with less than 60 months of survival to generate a more realistic survival proportion as their outcomes after 60 months are currently unknown. Survival curves of Soria's classification show clear distinction between the three main classes (Luminals, Basals and HER2) and their subclasses, showing strong association between survival outcomes with the classes. This was previously not reported in [16]. At the 5-Year survival time, we see the distinction in the three main classes in ssFCM-E-KKZ but not in ssFCM-M-KKZ. HER2, known to be associated with poor overall survival, is indicated in class 6 in all survival curves. Survival curve of triple-negative patients in classes 4 and 5, which belongs to the basal group, showed poorer survival than classes 1-3 in all survival curves. We feel that due to the very small number of patients being classified as class 4 out of the 413 patients, a more optimistic class 4 survival curve may have been drawn. Nevertheless, the distinction between the survival curves of the three main classes, evident in results of Soria's classification and ssFCM-E-KKZ indicates associations between the classes and survival which further supports Soria's classification.

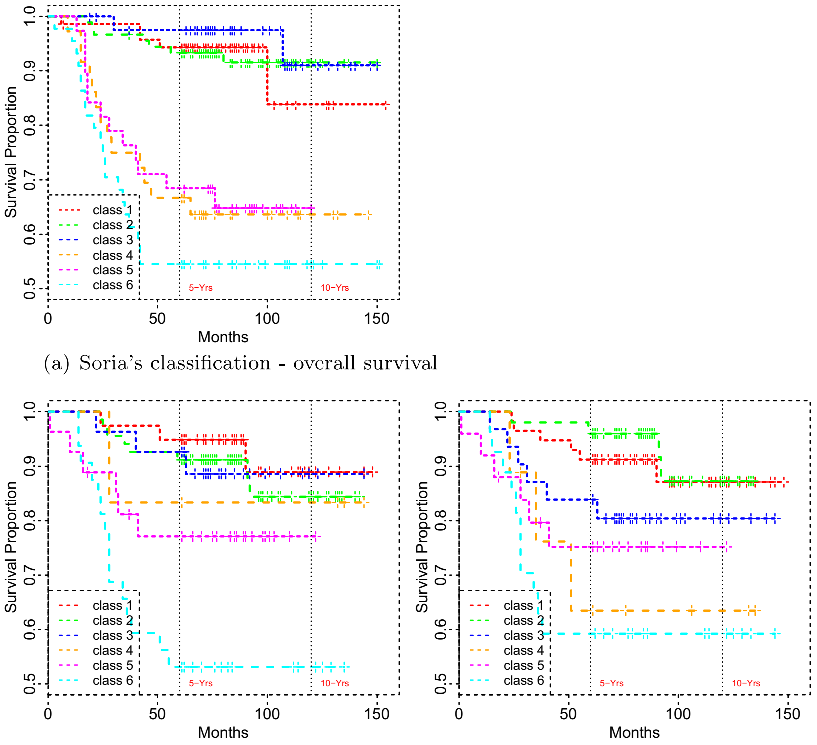
Table 8 Class distribution based on clinical parameters with ssFCM-E-KKZ and in brackets, ssFCM-M-KKZ.

	class 1	class 2	class 3	class 4	class 5	class 6	ϕ
Age							0.22 (0.23)
≤35	1 (5)	3 (1)	6 (4)	1 (1)	3 (3)	3 (3)	
35<Age≤45	56 (72)	62 (43)	33 (40)	6 (8)	20 (20)	32 (26)	
45<Age≤55	14 (15)	14 (13)	11 (11)	3 (4)	15 (18)	14 (10)	
>55	20 (32)	34 (26)	16 (14)	5 (9)	20 (17)	21(18)	
Total	91 (124)	113 (83)	66 (69)	15 (22)	58 (58)	70 (57)	
Grade							0.55 (0.54)
1	16 (22)	29 (23)	7 (8)	0 (0)	1 (1)	1 (0)	
2	48 (59)	53 (41)	16 (18)	1 (1)	4 (6)	13 (10)	
3	27 (43)	31 (19)	43 (43)	14 (21)	53 (51)	56 (47)	
Total	91 (124)	113 (83)	66 (69)	15 (22)	58 (58)	70 (57)	
Size							0.26 (0.22)
≤1.5cm	40 (45)	39 (29)	14 (23)	5 (7)	18 (16)	17 (13)	
1.5cm<Size≤2cm	7 (13)	13 (9)	10 (8)	0 (1)	8 (8)	15 (14)	
2cm<Size≤2.5cm	21 (30)	30 (24)	25 (21)	3 (5)	17 (19)	16 (13)	
2.5cm<Size≤3cm	11 (24)	19 (12)	11 (10)	3 (4)	11 (11)	14 (8)	
<3cm	12 (12)	12 (9)	6 (7)	4 (5)	4 (4)	8 (9)	
Total	91 (124)	113 (83)	66 (69)	15 (22)	58 (58)	70 (57)	
Lymph node stage							0.26 (0.19)
1	52 (71)	74 (55)	28 (35)	7 (13)	42 (38)	34 (25)	
2	36 (45)	27 (20)	32 (29)	5 (5)	14 (15)	25 (25)	
3	3 (8)	12 (8)	6 (5)	3 (4)	2 (5)	10 (6)	
Total	91 (124)	113 (83)	66 (69)	15 (22)	58 (58)	69 (56)	
Death							0.22 (0.18)
No	87 (116)	104 (79)	60 (60)	13 (17)	49 (50)	53 (44)	
Yes	3 (6)	8 (4)	3 (6)	1 (3)	6 (6)	15 (11)	
Total	90 (122)	112 (83)	63 (66)	14 (20)	55 (56)	68 (55)	
NPI							0.43 (0.44)
≤ 2.4 (EPG)	12 (17)	13 (9)	3 (4)	0 (0)	1 (0)	2 (1)	
2.4<NPI≤3.4 (GPG)	25 (29)	37 (34)	14 (15)	1 (1)	3 (3)	7 (5)	
3.4<NPI≤4.4 (MPG1)	21 (28)	34 (21)	11 (16)	3 (5)	23 (23)	16 (15)	
4.4<NPI≤5.4 (MPG2)	25 (32)	16 (10)	18 (19)	6 (11)	22 (22)	22 (15)	
<5.4(PPG)	8 (18)	13 (9)	20 (15)	5 (5)	9 (10)	23 (21)	
Total	91 (124)	113(83)	66 (69)	15 (22)	58 (58)	70 (57)	



(a) ssFCM-E-KKZ (b) ssFCM-M-KKZ

Fig. 4 Boxplots showing NPI distribution of the six classes obtained from ssFCM.



(a) Soria's classification - overall survival

(b) ssFCM-E-KKZ - overall survival (c) ssFCM-M-KKZ - overall survival

Fig. 5 Kaplan-Meier analysis of overall in different classes.

5 Discussion

We successfully reproduce Soria's classification with high agreement levels using ssFCM as a single method. To classify the remaining 413 patients, we use the techniques that reproduce Soria's classification with high agreement levels, that are ssFCM-E-KKZ and ssFCM-M-KKZ. Several different analysis are conducted to assess the classification. The distribution of biomarkers by class presented on the boxplots showed similar key characteristics of the six classes by Soria et al. The biplots showed the clusters generated from the classified patients' breast cancer type are located similarly to Soria's classification. Based on comparisons of boxplots of biomarkers and NPI values and biplots, ssFCM-E-KKZ produce classifications of patients (previously deemed not classified) that closely resemble Soria's classification. But, caution needs to be exercised in trying not to "force" the patients to belong to a class by choosing solutions from one technique. Instead, by using the confusion matrix and boxplot of NPI values from the two solutions, the patients belonging to mixed groups can be identified for further deeper analysis. A notable observation from the confusion matrix and NPI boxplot is that less disagreements are found in classes 4 and 5, allowing identification of possible triple negative patients. From the boxplots, we found that some biomarkers do not appear to contribute to class discrimination and may adversely affect classification accuracy. A further study to investigate in using feature selection to improve the classification results and identify important features is needed.

ssFCM-E-KKZ and ssFCM-M-KKZ were shown to produce high agreement with Soria's classification. We found that the fuzzy weights in the Mahalanobis distance did not produce a meaningful model for this dataset as much improved classification results were found using the non-fuzzy (original) Mahalanobis distance. Duda et al. [4] warns of the dangers of imposing structure instead of finding it when making a choice on distance metrics. Given a certain quantity of labelled data, ssFCM using Mahalanobis distance can perform competitively well as using Euclidean distance for this dataset. Previously in [16], the six clusters were found using Euclidean distance, which explains that ssFCM with Euclidean distance gave high classification accuracy. Had the dataset contain hyperellipsoidal clusters, ssFCM with Mahalanobis and fuzzy-Mahalanobis distances would most likely have performed better. We suspect that fuzzy Mahalanobis distance does not as work well with data patterns that form hyperspherical clusters as its very basis is to adapt to the shape and volume of the clusters. The fuzzy weights may an inaccurate adaptation of the shape and volume, particularly in a semi-supervised setting where the fuzzy weights of unlabelled data are equal for all clusters initially. We also incorporated initialisation techniques to achieve small but important improvements of classification results. Both ssFCM-E-KKZ and ssFCM-M-KKZ are capable of automatically identify the same six classes of breast cancer types as Soria's classification, with high degree of agreement with above 0.8 accuracy at least 20% of labelled data, which confirms Soria et al's six classes and addresses the issue of stability of their classification.

While all the features in NTBC are highly non-normally distributed, ssFCM has been able to produce very good classification results. In clustering techniques, Hair et al. stated that the requirement of normal distribution [7] has little effect. But more importantly, ssFCM is able to detect relevant areas of high concentration (see biplots in Figure 3 (b) and (c)) that are of importance using some labelled data, irrespective of the distribution.

Our examination of the classification of the remaining 413 patients interestingly reveals similar distribution of NPI values by class as study in [16] were found, which supports earlier claims of NPI providing discriminant information. Despite that these patients previously belonged to mixed classes [16], these new classification of the 413 patients showed characteristics consistent with those by Soria et al. Furthermore, the distinction between the three main classes found in the survival analysis of the classified breast cancer types for the 413 patients using ssFCM-E-KKZ, not only shows an association between the survival and breast cancer types but also supports Soria's classification. In the analysis of the biomarkers, their distinct values (shown in the cluster centres) and distributions (comparisons of boxplots with Soria's classification) verify their importance in characterising the classes and discriminating between them. The analysis with clinical information such as age, grade, NPI and survival showed significant associations between the classes already identified with these clinical information, which can help support a more accurate prognosis.

6 Conclusion

In this work, we have successfully identified the same breast cancer classes as those previously found and with a high level of agreement. We also classified the remaining 413 patients, previously unclassified, and showed that they exhibit similar characteristics as those in the already found six classes. Using both Euclidean distance and Mahalanobis distance, ssFCM was able to classify the 663 patients with high level of agreenent as Soria's classification. Small improvements can be further achieved using initialisation techniques. Based on further analysis of the classification of the 413 patients and their clinical parameters, the class characteristics found (based on clinical parameters) are consistent with those reported by Soria et al. [16]. In classifying the breast cancer types of the remaining 413 patients, we hope that we have provided a more accurate model for the prediction of breast cancer types for new patients, using both classification from Soria et al. and classification of the remaining 413 patients, which previously belonged to a mixture of classes, that can help support decision making. In this respect, our contribution is from a clinical point of view though application of computational techniques. We shall further our studies in using feature selection techniques to identify relevant (important) features that may improve classification results.

References

1. Abd El-Rehim, D.M., Ball, G., Pinder, S.E., Rakha, E., Paish, C., Robertson, J.F., Macmillan, D., Blamey, R.W., Ellis, I.O.: 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. International Journal of Cancer **116**(3), 340–350 (2005)
2. Al-Daoud, M.B., Roberts, S.A.: New methods for the initialisation of clusters. Pattern Recogn. Lett. **17**, 451–455 (1996)
3. Chiu, S.: Fuzzy Model Identification based on cluster estimation. Journal of Intelligent Fuzzy Systems **2**, 267–278 (1994)
4. Duda, R., Hart, P., Stork, D.: Pattern classification, 2 edn. Pattern Classification and Scene Analysis: Pattern Classification. Wiley-Interscience (2000)

5. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Some methods for classification and analysis of multivariate observations. Proc. Natl. Acad. Sci. USA (25), 14,863–14,868 (1998)
6. Gustafson, D.E., Kessel, W.C.: Fuzzy clustering with a fuzzy covariance matrix. In: IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes, vol. 17, pp. 761–766 (1978)
7. Hair, J., Black, W., Babin, B., Anderson, R.: Multivariate data analysis: A global perspective, 7 edn. Pearson Education (2010)
8. He, J., Lan, M., Tan, C.L., Sung, S.Y., Low, H.B.: Initialization of cluster refinement algorithms: a review and comparative study. In: Proceedings of 2004 IEEE International Joint Conference on Neural Networks, vol. 1, p. 3302 (2004). DOI 10.1109/IJCNN.2004.1379917
9. Katsavounidis, I., Jay Kuo, C.C., Zhang, Z.: A new initialization technique for generalized lloyd iteration. Signal Processing Letters, IEEE **1**(10), 144–146 (1994). DOI 10.1109/97.329844
10. Li, C., Liu, L., Jiang, W.: Objective function of semi-supervised fuzzy c-means clustering algorithm. In: IEEE International Conference on Industrial Informatics, pp. 737–742 (2008)
11. Maesschalck, R.D., Jouan-Rimbaud, D., Massart, D.: The mahalanobis distance. Chemometrics and Intelligent Laboratory Systems **50**(1), 1–18 (2000). DOI 10.1016/S0169-7439(99)00047-7
12. Makretsov, N., Huntsman, D., Nielsen, T., Yorida, E., Peacock, M., Cheang, M., Dunn, S., Hayes, M., van de Rijn, M., Bajdik, C., Gilks, C.: Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. Clinical Cancer Research **10**(18), 6143–6151 (2004)
13. Pedrycz, W., Waletzky, J.: Fuzzy clustering with partial supervision. IEEE Transactions on Systems, Man and Cybernetics **27**(5), 787–795 (1997)
14. Perou, C., Sørlie, T., Eisen, M., Van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S., Lønning, P., Børresen-Dale, A., Brown, P., Botstein, D.: Molecular portraits of human breast tumours. Nature **406**, 747–752 (2000)
15. Soria, D., Garibaldi, J., Biganzoli, E., Ellis, I.: A comparison of three different methods for classification of breast cancer data. In: Seventh International Conference on Machine Learning and Applications, 2008. ICMLA '08., pp. 619–624 (2008)
16. Soria, D., Garibaldi, J.M., Ambrogi, F., Green, A.R., Powe, D., Rakha, E., Macmillan, R.D., Blamey, R.W., Ball, G., Lisboa, P.J., Etchells, T.A., Boracchi, P., Biganzoli, E., Ellis, I.O.: A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. Computers in Biology and Medicine **40**(3), 318–330 (2010)
17. Sørlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Eystein Lønning, P., Børresen-Dale, A.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA **98**(19), 10,869–10,874 (2001)
18. Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geister, S., Demeter, J., Perou, C.M., Lønning, P.E., Brown, P.O., Børresen-Dale, A., Botstein, D.: Repeated observation of breast tumor subtypes in indepedent gene expression data sets. Proc Natl Acad Sci USA **100**(14), 8418–8423 (2003)
19. Stutz, C., Runkler, T.A.: Classification and prediction of road traffic using application-specific fuzzy clustering. IEEE Transactions on Fuzzy Systems **10**(3), 297–308 (2002)
20. Tari, L., Baral, C., Kim, S.: Fuzzy c-means clustering with prior biological knowledge. Journal of Biomedical Informatics **42**(1), 74–81 (2009)
21. Tou, J., Gonzales, R.: Pattern Recognition Principles. Addison-Wesley, Reading, MA (1974)