2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Identifying Stable Breast Cancer Subgroups Using Semi-supervised Fuzzy c-means on a Reduced Panel of Biomarkers

Daphne Teck Ching Lai and Jonathan M. Garibaldi

Abstract— **The aim of this work is to identify clinically-useful and stable breast cancer subgroups using a reduced panel of biomarkers. First, we investigate the stability of subgroups generated using two different reduced panels of biomarkers on clustering of breast cancer data. The stability of the subgroups found are assessed based on comparison of agreement levels using Cohen's Kappa Index on clustering solutions from ssFCM methodologies, consensus K-means and model-based clustering. The clustering solutions obtained from the feature set which achieve the higher agreement is chosen for further biological and clinical evaluation to establish the subgroups are clinically-useful. Using a ssFCM methodology, we identified seven clinically-useful and stable breast cancer subgroups using a reduced panel by Soria *et al.* So far, the stability of the subgroups identified using the reduced panel of biomarkers have not yet been investigated.**
Keywords: semi-supervised FCM, feature reduction, breast cancer classification, cluster stability

I. Introduction

Six clinically useful subgroups was found in the Nottingham Tenovus Breast Cancer (NTBC) dataset [1], which contains 1076 breast cancer cases with cellular information based on 25 protein biomarkers. The six classes were derived by reaching a consensus based on a semi-manual technique [1] featuring manually-generated rules and clinicians' expertise to combine solutions from several different clustering algorithms. As a result, 663 out of 1076 patients are classified. So far, no single clustering algorithm has been found to automatically identify the same subgroups. The overarching aim of our work is to 'reproduce' classification by Soria et al. [1] (Soria's classification for short) with a single clustering method, using all 1076 patient data. Previously, semi-supervised Fuzzy c-Means (ssFCM) has been shown to achieve this [2], [3], [4]. However, most of our studies have been applied on the 663 patients data where results are matched with Soria's classification for evaluation.

In this paper, we investigate the effect of two reduced panel of biomarkers on clustering of all 1076 patient data. Previously, 15 important features were identified using ssFCM and Naive Bayes-Recursive Feature Elimination (NB-RFE) [3]. In [5], however, 10 important features have been found using an exhaustive search of the best combination based on the Naive Bayes (NB) classification results. The same 10 important features were also used in to identify the key clinical phenotypes of breast cancer [6]. Also, these features are used in the generation of linguistic rule set a using fuzzy rule induction algorithm for breast cancer classification [7]. The existence of two different reduced panel of biomarkers (feature set) prompt the question as to which feature set will produce better subgroups, in terms of stability, biological meaningfulness and clinical relevance. In line with another previous study [3], our long term goal is to produce a clinically useful classification using fewer features (biomarkers), reducing the time and cost of running complex and expensive clinical tests.

While it makes sense that having more features give more discriminating power to distinguish between classes but, in practice, more features not only increase time requirements but also the irrelevant or redundant features can worsen classification accuracy. Hall [8] has described these features as "harmful redundancies". In our previous work [3], we have shown that higher agreement with Soria's classification can be achieved using feature selection.

Jain [9] raised a fundamental issue of clustering, which is the consistency of solutions from different clustering algorithms, that is, the stability of these different clustering solutions. Solutions from different clustering algorithms that are more consistent (stable) build higher confidence in the clusters found. For clustering of biomedical data, Bair and Tibshirani [10] explained the difficulty and importance of finding relevance between biological subgroups and clinical parameters for accurate prognosis. As unsupervised approaches often do not use clinical data to find subgroups, there is no assurance that the subgroups found would be related to the clinical outcome. Furthermore, the subgroups identified from clinical data may not be biologically meaningful. Statistical tests are, therefore, needed to determine the relevance between the biological subgroups and clinical data, which in turn can validate the subgroups identified. Therefore, to validate the biomedical subgroups, the subgroups have to be biologically meaningful and reproducible. Furthermore, the relevance between the subgroups and clinical parameters have to exist and the stability of these subgroups has to be addressed.

To do this, the breast cancer dataset are clustered with the two considered features sets from [3] and [5] using ssFCM, consensus K-means (CKM) and model-based clustering via Bayesian Information Criteria (MBIC). The feature sets are evaluated based on the stability of the subgroups produced. The stability are assessed based on agreement levels between clustering solutions using Cohen's Kappa index κ.

The aim is to identify the same subgroups as identified by Soria *et al.* [1] using a relevant reduced feature set in the breast cancer dataset and using all 1076 patients data. This means that the same relevant feature set is able to produce stable subgroups using different clustering algorithms. This investigation helps to identify stable subgroups and demonstrate whether these stable subgroups are biologically useful and clinically relevant. This investigation also helps to ascertain whether the reduced feature set can reproduce the same subgroups as with all 25 features.

Daphne Teck Ching Lai and Jonathan M. Garibaldi are with the School of Computer Science, University of Nottingham, United Kingdom (email: {dtl, jmg}@cs.nott.ac.uk).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

Algorithm 1 Semi-supervised fuzzy c-means [11]

1: Initialise c, labelled data membership matrix $\mathbf{F}$ and initial membership matrix $\mathbf{U}^{(0)}$
2: Calculate cluster centres using
$$\mathbf{v}_i = \frac{\sum_{j=1}^{N} u_{ij}^2 \mathbf{x}_j}{\sum_{k=1}^{N} u_{ij}^2}, \ 1 \leq i \leq c. \quad (2)$$
3: Compute fuzzy covariance matrices.
4: Compute squared distances d_{ij}^2 between cluster centres $\mathbf{v}_i$ and data patterns $\mathbf{x}_j$.
5: Update partition matrix, $\mathbf{U}$ using equation :
$$u_{ij} = \frac{1}{1+\alpha} \left\{ \frac{1 + \alpha(1 - b_j \sum_{l=1}^{c} f_{lj})}{\sum_{l=1}^{c} \left(\frac{d_{ij}}{d_{lj}}\right)^2} + \alpha f_{ij} b_j \right\} \quad (3)$$
6: If $||\mathbf{U}' - \mathbf{U}|| < \epsilon$, stop. Else, go to Line 2 with $\mathbf{U} = \mathbf{U}'$

II. Selected algorithms

A. Semi-supervised Fuzzy c-Means

The objective function of the ssFCM proposed by Pedrycz and Waletzky [11] contains unsupervised learning in the first term and supervised learning in the second term as follows:

$$J = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^p d_{ij}^2 + \alpha \sum_{i=1}^{c} \sum_{j=1}^{N} (u_{ij} - f_{ij} b_j)^p d_{ij}^2, \quad (1)$$

where u_{ij} is the membership value of data pattern j in cluster i, c is the number of clusters, d_{ij} the distance (Euclidean in this case) between data pattern j and cluster centre v_i, f_{ij} the membership value of labelled data pattern j in cluster i, b_j indicates if data pattern j is labelled, 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 such that supervised learning does not dominate. The authors recommend α to be proportional to N/M, where M is the number of labelled data. The algorithm is summarised in Algorithm 1.

The ssFCM algorithm can be enhanced using initialisation techniques [12] to calculate initial clusters, which we previously used for ssFCM classification [2]. Another enhancement is the adjustment of scaling parameter α to adjust the influence of labelled data [13].

B. Consensus k-means

A simple algorithm is devised, which is referred as CKM for short, to reach a consensus of K-means clustering solutions as shown on Algorithm 2.

The parameter ϵ is chosen by visual inspection of biplot of the six lists, which are essentially the six clusters. The choice of ϵ is arbitrary. A small ϵ will increase the tendency for clusters to merge or overlap and a large ϵ will create more compact clusters at the cost of ignoring some data patterns.

C. Model-based clustering via Bayesian information criterion

Fraley and Raftery [14] implemented a model-based clustering (MBIC) which uses the Maximum *A Posteriori* (MAP) estimate from a Bayesian analysis to estimate model parameters, instead of Maximum Likelihood Estimation (MLE) in the Expectation-Maximization (EM) algorithm and a modified Bayesian Information Criterion (BIC) for model selection. MAP

Algorithm 2 Consensus k-means

1: Run K-means 5000 times to generate c (six) clusters. The output is a 5000 $\times$ N matrix containing cluster labels $c_1, ...c_N$ for data patterns $x_1, ..., x_N$ for each run. A large number of run is chosen to ensure that similar data can be identified over many runs.
2: For data pattern x_i, count the number of times it is in the same cluster as other data patterns, $count_{ij} = count(c_i == c_j)$, in all runs.
3: **repeat**
4: For data pattern x_i, a list, l_i for $i = 1, ..., N$ is created containing other data patterns that share the same clusters for $count_{ij} > \epsilon$.
5: If $x_j \in l_i$ and $length(l_i \cap l_j) > 20$, each list l_i is then updated by performing a union with all other lists that share common data patterns $l_i = l_i \cup l_j, j \neq i$. If the other list l_j fulfills this condition with l_i, l_j is deleted.
6: The largest six lists are chosen as the clusters and other lists, usually with much smaller number of members are ignored.
7: Present the six lists in a biplot.
8: **until** A biplot of clusters most similar to those by Soria *et al.* is produced. Repeat with different ϵ values if necessary. [1]

is used to avoid the failure of the EM algorithm in the presence of singularities or degeneracies. The mixture model with density for generating data $y = (y_1, ..., y_n)$ in model-based clustering is defined as:

$$f(y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k), \quad (4)$$

where $f_k(y_i|\theta_k)$ is a probability distribution with parameters θ_k, τ_k is the probability of belonging to the k^{th} component and $\theta_k = (\mu_k, \Sigma_k)$, μ_k are the means and Σ_k the covariances of f_k. These parameters of the model are estimated using MLE in the EM algorithm.

To eliminate EM failure to converge due to singularity in covariance estimate, the authors proposed a prior distribution on the parameters that can eliminate the singularity problem while maintaining stability on results obtainable without a prior probability. The Bayesian predictive density for the data is in the form:

$$\mathcal{L}(Y|\tau_k, \mu_k, \Sigma_k)\mathcal{P}(\tau_k, \mu_k, \Sigma_k|\theta), \quad (5)$$

where $\mathcal{L}$ is the mixture likelihood:

$$\begin{aligned}\mathcal{L}(Y|\tau_k, \mu_k, \Sigma_k) &= \prod_{j=1}^{n} \sum_{k=1}^{G} \tau_k \phi(y_i|\mu_k, \Sigma_k) \\ &= \prod_{j=1}^{n} \sum_{k=1}^{G} \tau_k |2\pi\Sigma_k|^{-\frac{1}{2}} \\ &\quad \exp\left\{ -\frac{1}{2}(y_j - \mu_k)^T \Sigma_k^{-1} (y_j - \mu_k) \right\},\end{aligned}$$

and $\mathcal{P}$ is a prior distribution on the parameters τ_k, μ_k, σ_k and θ. The f_k in (4) is the multivariate Gaussian density ϕ with parameters μ_k as its mean and Σ_k as its covariance.

The BIC [15] selects the best fitted model from a finite set of models using maximum likelihood. It is defined [14] as:

$$BIC \equiv 2\log\mathcal{L}_{max} - k\log(N) \quad (6)$$

where $\mathcal{L}_{max}$ is the maximum likelihood of the estimated model, k the number of parameters in the model and N the number of data patterns used in the estimation. The BIC is modified by replacing the first term in (6), $2\log\mathcal{L}_{max}$ by twice the log-likelihood evaluated using MAP in (5).

III. The Nottingham Tenovus Breast Cancer dataset

The Nottingham Tenovus Breast Cancer (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 scored 0 to 3, which correspond to negative, weak, moderate and strong positivity. The H-score ranges between 0 and 300, based on the formula:

$$\begin{aligned} \text{H-score} &= (1 \times \% \text{ of cells with intensity } 1) \\ &+ (2 \times \% \text{ of cells with intensity } 2) \\ &+ (3 \times \% \text{ of cells with intensity } 3) \end{aligned} \quad (7)$$

The 25 protein biomarkers (features) are the same ones listed in [1] and illustrated on Table I. The dataset also contains clinical data such as histologic grade, histologic tumour type, vascular invasion, tumour size, lymph node stage, patient age and menopausal status. Survival (in months) from the date of primary treatment to the time of death is recorded at 3-months intervals initially, then every 6 months, and finally, annually for a range of 1-192 months, with a median period of 58 months. The Nottingham Prognostic Index (NPI) [16] score is also recorded. It is calculated based on prognostic factors according to the formula: NPI Score $= (0.2 \times$ tumour size$) +$ histologic grade $+$ lymph node stage where a poor prognosis is indicated by a high NPI score.

In Soria's classification [1], 663 data patterns are classified while 413 remains not classified (n.c) , as shown in Table II. Based on classification by Soria *et al.* [1], there are three main clinical groups, Luminal, Basal and HER2. These main groups are further divided into six subgroups where class 1, 2 and 3 belong to the Luminal group. Class 4 and 5 belong to the Basal group and class 6 to HER2. Each class is named (in square brackets) and their key features identified by Soria *et al.* [1] are tabulated in Table III. In [7], Soria *et al.* proposed a quantifier-based classification system with a reduced panel of biomarkers to refine the previous classification in [1], classifying the entire dataset into seven subgroups where class 6 is split into two classes, class 6 (HER+/ER+) and class 7 (HER+/ER-).

IV. Experimental methods

First, using ssFCM methodologies, CKM and MBIC, the dataset is clustered into six subgroups where the sixth subgroup (HER2) is manually split into two subgroups (HER2/ER+) and (HER2/ER-) [6]. For ssFCM, 663 labels from Soria's classification [1] is used as these have been shown to be clinically-useful. In our previous study [4], we experimented using FCM, Hierarchical Clustering (HC), K-means and MBIC and found that the latter two produced favourable results. For this reason, the two algorithms are used in this study.

Next, the stability of the clustering solutions based on each of the two reduced feature sets are assessed. The stability between clustering solutions are calculated using the Cohen's κ Index which is available using the command `confusionMatrix` from the `caret` R package [17]. The ssFCM methodologies that are explored are ssFCM, ssFCM with KKZ initialisation proposed by Katsavounidis, Kuo and Zhang [12] (denoted as SK) and ssFCM with KKZ and α set to 30 (denoted as SKA). The enhanced ssFCM methodologies SK [3] and SKA (unpublished) were previously found to improve ssFCM performance. They all use Euclidean distance. Furthermore, ssFCM clustering solution with 25 features (ssFCM25) are also compared as a clustering solution that maintain a high agreement with all

TABLE I

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

* 10 features identified by Soria *et al.* [5]
+ 15 features identified by Lai and Garibaldi [3]

TABLE II

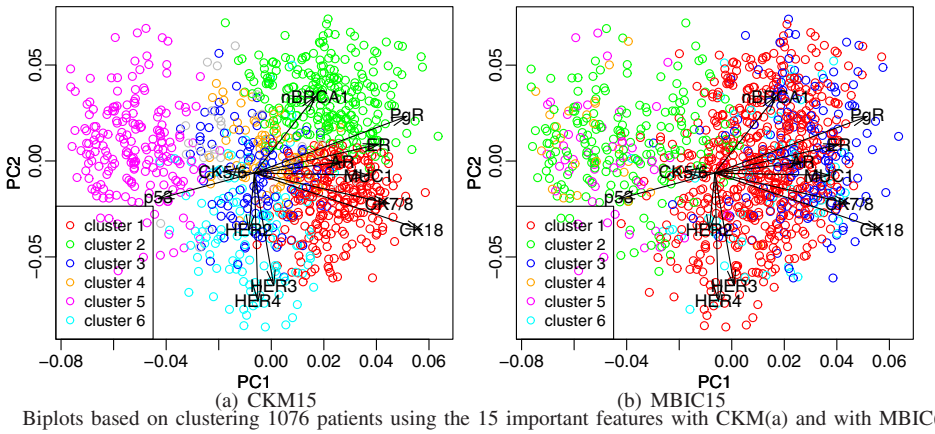
Population of each class and the number of not classified (n.c) and classified (c) data patterns according to classification by Soria *et al.* [1]

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

TABLE III

Key biomarkers for the 6 classes (c) [1].

c	Subgroup Name	Biomarkers
1	Luminal A	ER+ PgR+ CK7/8+ CK18+ CK19+ HER3+ HER4+
2	Luminal N	ER+ PgR+ CK7/8+ CK18+ CK19+ HER3- HER4-
3	Luminal B	ER+ PgR- CK7/8+ CK18+ CK19+ HER3+ HER4+
4	Basal-p53 altered	ER- p53+ CK5/6+ CK14+
5	Basal-p53 normal	ER- p53- CK5/6+ CK14+
6	HER2	ER- HER2+



(a) CKM15 (b) MBIC15

Fig. 1. Biplots based on clustering 1076 patients using the 15 important features with CKM(a) and with MBIC(b).

solutions using as few features as possible is desirable.

Confusion matrices based on the highest agreement levels for each clustering algorithm and reduced feature set are shown. To show where the disagreement occurs with respect to individual subgroups (classes), sensitivity and specificity measures are used. Sensitivity measures the rate of true positives and specificity measures the rate of true negative. The confusion matrix, Cohen's κ Index and sensitivity and specificity measures are implemented using `confusionMatrix` from the `caret` R package [17]. To ensure that the subgroups identified are biologically useful and clinically relevant, biological (using biplots) and clinical (using association measures) evaluation are conducted on the clustering solutions. To measure association between subgroups with clinical parameters, the Cramer V coefficient is used. The p-value presented in brackets is based on Pearson's chi-squared test of independence. This is implemented using the `assocstats` function in the `vcd` R package [18].

V. Results

Table IV shows the agreement levels between the various clustering methods using the 15 features identified in [3] and 10 features in [5]. Table V shows the agreement between the different clustering solutions using 10 and the original 25 features. The reduced feature set used is indicated with the number of features at the end of the clustering method used. Figure 1 shows that CKM15 and MBIC15 produced different subgroups from those identified by Soria *et al.* [1] such that the cluster labels are difficult to align for direct comparison, particularly for MBIC15. Thus, only biplots are presented for them. Although CKM15 has high agreement with ssFCM25, the biplot showed that CKM15 cannot distinguish between class 4 and class 5. High agreement of above 0.87 is maintained between ssFCM methodologies with 10 features and ssFCM25. Agreement with ssFCM25 is higher for SKA10 than SK10 and ssFCM10. These observations indicate that more stable subgroups can be generated using the 10 features [5] than using 25 or 15 features. For this reason, SKA10, CKM10 and MBIC10 are chosen for further analysis.

Furthermore, comparing Table V with Table 4 in [1], the agreement between clustering solutions of HC (agglomerative), ART, KM and PAM in [1] is lower (agreement range of 0.2 to 0.6) than the agreement between clustering solutions here obtained using the reduced feature set (agreement of above 0.6).

TABLE IV

Agreement levels using Cohen's κ Index of clustering solutions based on reduced panels of 15 and 10 features.

	6 subgroups			7 subgroups		
	CKM15	MBIC15	ssFCM25	CKM15	MBIC15	ssFCM25
ssFCM25	0.727	-	-	0.729	-	-
ssFCM15	0.730	-	0.983	0.732	-	0.983
SK15	0.728	-	0.976	0.730	-	0.976
SKA15	0.730	-	0.968	0.733	-	0.968
	CKM10	MBIC10	ssFCM25	CKM10	MBIC10	ssFCM25
ssFCM25	0.650	0.625	-	0.716	0.628	-
ssFCM10	0.715	0.691	0.873	0.716	0.694	0.874
SK10	0.717	0.691	0.874	0.719	0.694	0.875
SKA10	0.699	0.674	0.881	0.701	0.676	0.882

TABLE V

Comparing agreement (Cohen's κ Index) between clustering solutions using 10 [5] and 25 features.

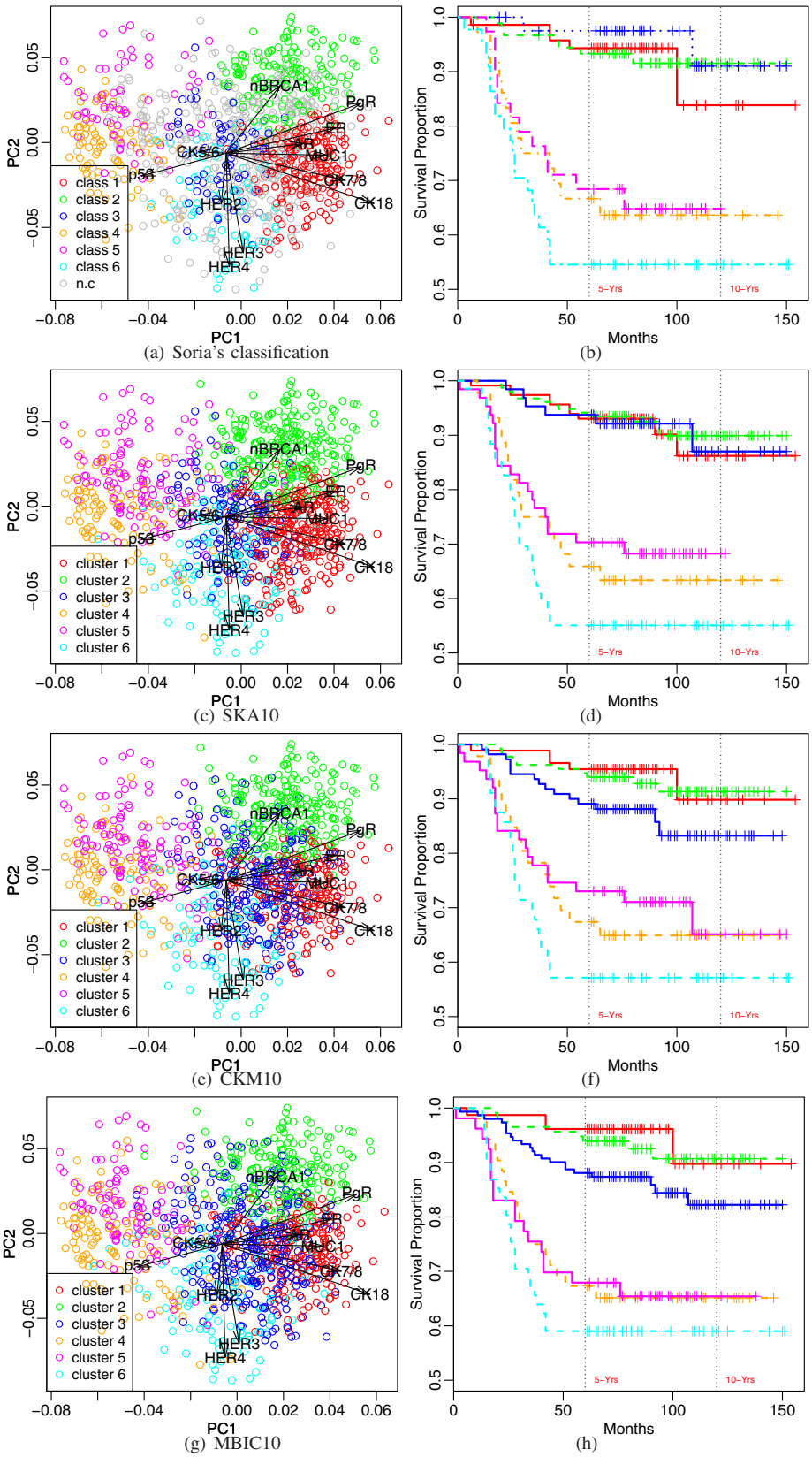
	6 subgroups		7 subgroups	
	CKM10	MBIC10	CKM10	MBIC10
SKA10	0.699	0.674	0.701	0.676
CKM10	-	0.860	-	0.861
	CKM25	MBIC25	CKM25	MBIC25
ssFCM25	0.587	0.765	0.600	0.767
CKM25	-	0.590	-	0.592

This further supports that more stable subgroups are obtained using the reduced panel of 10 features.

Comparison between ssFCM25 and SKA10 using confusion matrices in Table VI show that there is only small disagreements between the three main groups, and the disagreements within the same main groups between their respective subgroups are considered small. Based on the confusion matrices and the sensitivity and specificity measures, it is observed that disagreements tend to occur within clusters 1 or 3 where there is low sensitivity (in italics) found using CKM10 and MBIC10. Nevertheless, their average sensitivity of above 0.7 and specificity of above 0.9.

Based on Figure 2, identical subgroups (as shown on the biplots) as those from Soria's classification could be found using the three clustering algorithms with 10 features. Furthermore, their respective survival curves show clinical relevance in terms of overall survival outcome. The separability between the six survival curves which correspond to their biological subgroups reflected the three main breast cancer groups and six subgroups, similar to Soria's classification.

Figure 3 shows the survival curves based on subgroups



(a) Soria's classification (b)

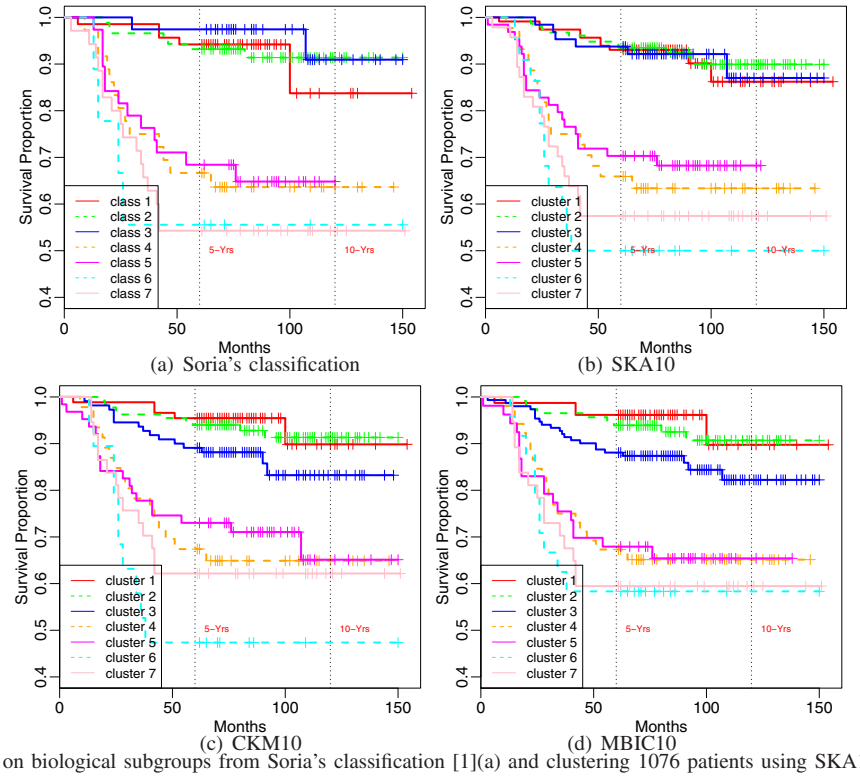
(c) SKA10 (d)

(e) CKM10 (f)

(g) MBIC10 (h)

Fig. 2. PCA biplots of Soria's classification [1](a) and clustering 1076 patients using SKA10(c), CKM10(e) and MBIC10(g) and their respective survival curves.

identified by the clustering algorithms with HER2 group divided into two. Survival curves based on subgroups by Soria's classification and MBIC10 distinctively maintain both the 3 main groups and their respective subgroups. For CKM10, survival curves for Basal and HER 2 group is not clear while for SKA10, the distinction between these two main groups is not as clear as Soria's or MBIC10's. Table VII shows the survival curve differences based on subgroups from SKA10 using G-rho family of tests proposed by Harrington and Fleming [19]. The test determines whether there is a difference between one or



(a) Soria's classification (b) SKA10

(c) CKM10 (d) MBIC10

Fig. 3. Survival curves based on biological subgroups from Soria's classification [1](a) and clustering 1076 patients using SKA10(b), CKM(c) and MBIC(d).

TABLE VI

CONFUSION MATRICES BETWEEN CLUSTERING SOLUTIONS FROM ssFCM25 AND SKA10, CKM AND MBIC.

	1	2	3	4	5	6	total
SKA10	ssFCM25						
1	275	12	7	0	0	8	302
2	7	241	4	0	4	4	260
3	15	5	125	0	3	3	151
4	0	1	2	92	0	6	101
5	0	4	1	1	119	5	130
6	4	2	1	2	2	121	132
Total	301	265	140	95	128	147	1076
Sensitivity	0.914	0.909	0.893	0.968	0.930	0.823	
Specificity	0.965	0.977	0.972	0.991	0.988	0.988	
P-value	0.00						
CKM10	ssFCM25						
1	186	2	33	0	1	1	223
2	13	208	8	0	4	0	233
3	98	49	68	0	15	25	255
4	0	1	1	87	3	5	97
5	0	1	22	4	94	2	123
6	2	1	2	2	2	102	111
o.c	2	3	6	2	9	12	34
Total	301	265	140	95	128	147	1076
Sensitivity	0.622	0.794	*0.507*	0.935	0.790	0.756	
Specificity	0.950	0.968	0.794	0.989	0.969	0.990	
P-value	0.00						
MBIC10	ssFCM25						
1	172	5	24	0	1	1	203
2	5	182	7	0	5	0	199
3	119	72	98	0	17	30	336
4	0	1	3	92	2	15	113
5	0	0	4	0	100	2	106
6	5	5	4	3	3	99	119
Total	301	265	140	95	128	147	1076
Sensitivity	*0.571*	0.687	0.700	0.968	0.781	0.673	
Specificity	0.960	0.979	0.746	0.979	0.994	0.978	
P-value	0.00						

more survival curves where a p-value of less than 0.05 means that they are different. It is implemented using the `survdiff` function from the `survival` R package [20]. Survival curves differ significantly between Luminals (clusters 1-3) and the other 2 groups, Basals (clusters 4 and 5) and HER2 (clusters 4 and 5), showing poorer prognosis in the more aggressive groups Basals and HER2 [1]. Furthermore, the separability between the survival curves which reflects the prognosis of biological subgroup indicates clinical relevance, which indicates that these subgroups are clinically useful.

TABLE VII

DIFFERENCES IN SURVIVAL CURVES IN SKA10 USING KAPLAN-MEIER P-VALUES.

Clusters	1	2	3	4	5	6
2	0.743					
3	0.989	0.785				
4	1.58E-05	6.15E-07	3.84E-04			
5	6.40E-05	4.00E-06	1.25E-03	0.632		
6	1.27E-08	1.80E-10	3.33E-06	0.260	0.105	
7	1.01E-07	1.16E-09	2.18E-05	0.466	0.214	0.575

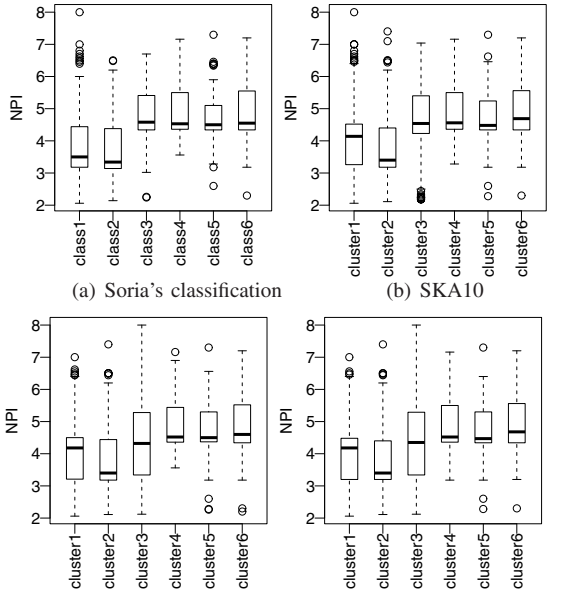
As the 10 features produced stable subgroups using different clustering algorithms and ssFCM is able to retain Soria's classification completely, clinical association of the subgroups found by SKA10 is presented in Table VIII. Based on Cramer's V and, presented in brackets, p-values, significant association with clinical parameters that are not involved in clustering has been found. Note that patients with missing clinical information (indicated by discrepancies in the total for each class) are not included in the Cramer's V test. The clinical association between six identified by other clustering methods are compared in Table IX. The subgroups identified by SKA10 have clinical associations that are competitive with Soria's classification and that are higher in more clinical parameters than CKM10 and MBIC10. Note that Soria's classification considers only 663 patients, while subgroups from SKA10, CKM10 and MBIC10 consider all 1076 patients.

TABLE VIII

CLINICAL ASSOCIATION OF SKA10 CLUSTERS. THE CRAMER'S V COEFFICIENT (IN ITALICS) AND P VALUES (IN BRACKETS) ARE SHOWN.

Parameter *Cramer's V*	cla.1	cla.2	cla.3	cla.4	cla.5	cla.6
Age 0.13 (0.00)						
≤35	12	6	5	12	8	5
35<Age≤45	150	123	92	30	47	53
45<Age≤55	50	41	20	27	35	29
>55	90	90	34	32	40	45
Total	302	260	151	101	130	132
Grade 0.42 (0.00)						
1	71	70	15	0	3	1
2	129	142	37	2	13	20
3	101	48	99	99	114	111
Total	301	260	151	101	130	132
Size 0.12 (0.00)						
≤1.5cm	110	109	42	19	31	29
1.5cm<Size≤2cm	31	19	21	9	24	28
2cm<Size≤2.5cm	78	71	43	28	35	37
2.5cm<Size≤3cm	52	44	25	24	27	23
<3cm	31	17	20	21	13	15
Total	302	260	151	101	130	132
Stage 0.14 (0.00)						
1	190	174	75	60	94	61
2	96	67	63	28	25	53
3	16	18	13	13	11	16
Total	302	259	151	101	130	130
Death 0.26 (0.00)						
No	283	244	139	81	105	99
Yes	10	13	6	16	20	31
Total	293	257	145	97	125	130
NPI 0.23 (0.00)						
≤ 2.4 (EPG)	51	47	9	0	1	3
2.4<NPI≤3.4 (GPG)	74	93	21	1	8	9
3.4<NPI≤4.4 (MPG1)	78	60	38	35	44	37
4.4<NPI≤5.4 (MPG2)	62	39	46	35	55	38
<5.4(PPG)	37	21	37	30	22	45
Total	302	260	151	101	130	132

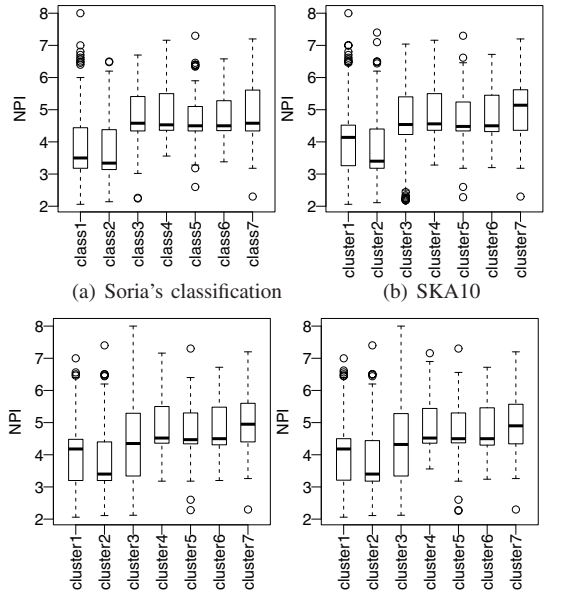
Figures 4 and 5 show the NPI distribution of the 6 and 7 subgroups based on the different clustering algorithms respectively. Subgroups from SKA10 produced similar NPI distributions as Soria's classification while subgroups from CKM10 and MBIC10 have similar NPI distributions where their cluster 3 have a higher NPI dispersion than those from Soria's classification and SKA10. Furthermore, similar NPI distributions for all seven subgroups as those in [6] were found using CKM10 and MBIC10.



(a) Soria's classification (b) SKA10

(c) CKM10 (d) MBIC10

Fig. 4. Boxplots showing NPI distribution for 6 subgroups.



(a) Soria's classification (b) SKA10

(c) CKM10 (d) MBIC10

Fig. 5. Boxplots showing NPI distribution for 7 subgroups.

VI. DISCUSSION

Using the 10 features identified in [5], stable subgroups have been found, which were assessed using agreement levels between three clustering algorithms. Based on the clinical evaluation using association between biological subgroups and clinical parameters, survival analysis and NPI boxplot analysis, the subgroups identified using all three clustering algorithms to be clinically relevant with ssFCM subgroups having highest association with grade, stage, NPI and death in comparison with subgroups identified by CKM and MBIC.

Further comparison were made with the 7 subgroups identified by Green *et al.* [6], the latest development of subgroup identification in the NTBC dataset. The HER2 group is manually split into two such that those with ER more than zero goes into the HER2/ER+ group (class/cluster 6) and those with zero ER expression goes into the HER2/ER- group (class/cluster 7). On comparison of clinical association with subgroups by Green *et al.* [6], competitive association levels were found with the subgroups found using the ssFCM framework. NPI distributions of subgroups based on CKM10 and MBIC10 are similar to those of Green *et al.* [6]. This further ascertain the importance of the 10-feature set in identifying stable subgroups using different clustering techniques and methodologies.

Based on the observation of increased agreement in clustering solutions between ssFCM and CKM and more drastic increased agreement between CKM and MBIC, this suggests that the subgroups from different clustering algorithms stabilise with a suitable, reduced feature set. The significant increase in agreement between CKM and MBIC with 10 features warrants further investigation between agreeing solutions, which may produce clearer distinction between Basal and HER2 group for the 7 subgroups of CKM10.

TABLE IX

CLINICAL ASSOCIATION OF CLUSTERING BASED ON SORIA *et al.* [1], SKA10, CKM10 AND MBIC10 MEASURED BY CRAMER'S V AND THEIR RESPECTIVE P-VALUES IN BRACKETS BASED ON 6 AND 7 SUBGROUPS (SG).

	Soria *et al.*[1]		SKA10		CKM10		MBIC10	
Parameters	6 SG	7 SG	6 SG	7 SG	6 SG	7 SG	6 SG	7 SG
Age	0.15 (0.00)	0.16 (0.00)	0.13 (0.00)	0.13 (0.00)	0.15 (0.00)	0.16 (0.00)	0.14 (0.00)	0.14 (0.00)
Grade	0.47 (0.00)	0.47 (0.00)	0.42 (0.00)	0.42 (0.00)	0.40 (0.00)	0.40 (0.00)	0.40 (0.00)	0.40 (0.00)
Size	0.15 (0.00)	0.15 (0.00)	0.12 (0.00)	0.12 (0.00)	0.13 (0.00)	0.13 (0.00)	0.12 (0.00)	0.12 (0.00)
Stage	0.15 (0.001)	0.16 (0.001)	0.14 (0.00)	0.14 (0.00)	0.10 (0.029)	0.10 (0.036)	0.11 (0.005)	0.11 (0.011)
NPI	0.26 (0.00)	0.26 (0.00)	0.23 (0.00)	0.24 (0.00)	0.22 (0.00)	0.22 (0.00)	0.21 (0.00)	0.22 (0.00)
Death	0.30 (0.00)	0.30 (0.00)	0.26 (0.00)	0.26 (0.00)	0.24 (0.00)	0.25 (0.00)	0.23 (0.00)	0.23 (0.00)

The 15 features found using ssFCM and NB-RFE are useful for achieving high classification accuracy when assigning new patients to classes [2], but from this study, they may not be useful for finding stable subgroups when used with unsupervised clustering algorithms. This may be due to several reasons. The 15 features were identified based on Soria's classification labels of the 663 patients [1], not all 1076 patients. Perhaps, we should derive features using ssFCM and NB-RFE [3] using Soria's latest classification [7]. Furthermore, Soria's classification may be incorrect, given only 663 out of 1076 (about 61% of 1076) were classified. This suggests Soria's subgroups may not be as robust as assumed although they are shown to be biologically meaningful. This can greatly affect the selection of the 15 features found and on the stability of subgroups, as we chose CKM subgroups that highly agree with Soria's subgroups [1]. Furthermore, other useful collections of protein biomarkers that can be rationalised by other expert clinicians may exist.

The increased stability of subgroups generated by clustering algorithms using a reduced panel of protein biomarkers opened up two research questions, which to the best of our knowledge, are currently not answered:

1) Can feature selection help clustering algorithms produce more stable clusters?
2) Can the stability of clusters be an evaluation criteria for unsupervised feature selection using clustering algorithms to find relevant features?

VII. SUMMARY

In this study, clustering is performed using ssFCM with experimentation on two different feature sets, 10 from [5] and 15 [3]. Using 15 features, ssFCM achieved high agreement with ssFCM25. But, poor agreement were found using CKM15 and MBIC15, indicating that the subgroups found were unstable. Using the 10 features, SKA10 identified stable breast cancer subgroups, which were assessed based on agreement measure with solutions from CKM10 and MBIC10. Furthermore, results from biological and clinical evaluation showed that the subgroups are biologically useful and clinically relevant. Therefore, stable and clinically useful breast cancer subgroups have been successfully identified using a single clustering method, SKA10. The six subgroups found using 10 features were manually split into seven to make comparison with other subgroups found in [6]. Competitive clinical association and similar NPI distributions in the seven subgroups were found, further confirming the importance of the 10 features in identifying stable subgroups.

REFERENCES

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

[2] D. T. C. Lai, J. M. Garibaldi, D. Soria, and C. M. Roadknight, "A methodology for automatic classification of breast cancer immunohistochemical data using semi-supervised fuzzy c-means," *Central European Journal of Operations Research*, vol. (in press), pp. 1–25, 2013.

[3] D. T. C. Lai and J. M. Garibaldi, "Improving semi-supervised fuzzy c-means classification of breast cancer data using feature selection," in *2013 IEEE International Conference on Fuzzy Systems*, 2013, pp. 1–8.

[4] ——, "A preliminary study on automatic breast cancer data classification using semi-supervised fuzzy c-means," *International Journal of Biomedical Engineering and Technology*, vol. 13, no. 4, pp. 303–322, 2013.

[5] D. Soria, J. Garibaldi, E. Biganzoli, and I. Ellis, "A comparison of three different methods for classification of breast cancer data," in *Seventh International Conference on Machine Learning and Applications*, 2008, pp. 619–624.

[6] A. Green, D. Powe, E. Rakha, D. Soria, C. Lemetre, C. Nolan, F. Barros, R. Macmillan, J. Garibaldi, G. Ball, and I. Ellis, "Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers," *British Journal of Cancer*, vol. 109, pp. 1886–1894, 2013.

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

[8] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[9] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[10] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLoS Biology*, vol. 2, no. 4, p. e108, 2004.

[11] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 27, no. 5, pp. 787–795, May 1997.

[12] I. Katsavounidis, C.-C. Jay Kuo, and Z. Zhang, "A new initialization technique for generalized lloyd iteration," *Signal Processing Letters, IEEE*, vol. 1, no. 10, pp. 144–146, October 1994.

[13] A. Bouchachia and W. Pedrycz, "Enhancement of fuzzy clustering by mechanisms of partial supervision," *Fuzzy Sets and Systems*, vol. 157, no. 13, pp. 1733–1759, 2006.

[14] C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of Classification*, vol. 24, no. 2, pp. 155–181, 2007.

[15] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[16] M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis, "The nottingham prognostic index in primary breast cancer," *Breast cancer research and treatment*, vol. 22, no. 3, pp. 207–219, 1992.

[17] M. Kuhn, "Variable selection using the `caret` package," 2011 Last assessed: 5th November 2013. [Online]. Available: http://cran.r-project.org/web/packages/caret/vignettes/caretSelection.pdf

[18] D. Meyer, A. Zeileis, K. Hornik, F. Gerber, and M. Friendly, "The `vcd` package," 2013 Last assessed: 13th December 2013. [Online]. Available: http://cran.r-project.org/web/packages/vcd/vcd.pdf

[19] D. P. Harrington and T. R. Fleming, "A class of rank test procedures for censored survival data," *Biometrika*, vol. 69, no. 3, pp. 553–566, 1982.

[20] T. Therneau, "The `survival` package," 2013 Last assessed: 13th December 2013. [Online]. Available: http://cran.r-project.org/web/packages/survival/survival.pdf