A Novel Framework to Elucidate Core Classes in a Dataset

Daniele Soria, *Member, IEEE* and Jonathan M. Garibaldi, *Member, IEEE*

***Abstract*— In this paper we present an original framework to extract representative groups from a dataset, and we validate it over a novel case study. The framework specifies the application of different clustering algorithms, then several statistical and visualisation techniques are used to characterise the results, and core classes are defined by consensus clustering. Classes may be verified using supervised classification algorithms to obtain a set of rules which may be useful for new data points in the future. This framework is validated over a novel set of histone markers for breast cancer patients. From a technical perspective, the resultant classes are well separated and characterised by low, medium and high levels of biological markers. Clinically, the groups appear to distinguish patients with poor overall survival from those with low grading score and better survival. Overall, this framework offers a promising methodology for elucidating core consensus groups from data.**

I. Introduction

Clustering has become a widely used approach to extrapolate important information from data and to separate different groups that share similar characteristics within them. Cluster analysis may be thought of as the discovery of distinct and non-overlapping sub-partitions within a larger population [1]. Many different clustering techniques are known today, but often only a few selected methods are used in any given domain. For example, in breast cancer studies, researchers tend to focus on a single algorithm, usually hierarchical clustering [2], [3], [4]. Choosing which method to use is not an easy task, as different clustering techniques return different groupings. Consequently, it has been demonstrated [5], [6] that the use of several methods is preferable in order to extract as much information as possible from the data.

When using more than one algorithm, it is then common to define a consensus across the results [7] in order to integrate diverse sources of similarly clustered data [8] and to deal with the stability of the results obtained from different techniques. Several approaches have been proposed for this task. Kellam and colleagues [7] identified robust clusters by the implementation of a new algorithm called 'Clusterfusion'. It takes the results of different clustering algorithms and generates a set of robust clusters based upon the consensus of the different results of each algorithm.

Another approach, suggested by Monti and colleagues [1], deals with class discovery and clustering validation tailored to the task of analysing gene expression data. The methodology, termed 'consensus clustering', provides a method, in conjunction with resampling techniques, to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the discovered clusters.

Filkov and Skiena suggested to exploit the popularity of cluster analysis of biological data by integrating clusterings from existing data sets into a single representative clustering based on pairwise similarities of the clusterings. The goal of their consensus clustering was to eliminate the likely noise and incongruencies from the original classifications. Their proposed representative clustering was the one that minimised the distance to all the other partitions [8].

In another approach, Swift and colleagues used consensus clustering to improve confidence in gene-expression analysis, on the assumption that microarray analysis using clustering algorithms can suffer from lack of inter-method consistency in assigning related gene-expression profiles to clusters [9]. To assess gene-expression cluster consistency, the use of the weighted-kappa metric was analysed. This metric is generally used as a comparison between two data partitions as it rates the agreement between the classification decisions made by two or more observers. In this approach, the two observers are the clustering methods.

In addition to clustering methods, supervised classification techniques are widely used to learn classification rules from a set of labelled cases (training set) to label new cases in a test set. Many different supervised classification methods have been developed in recent years, such as Neural Networks, Classification Trees, Bayesian Classifiers and many more.

In this paper, an original algorithmic framework to elucidate a set of core groups in a dataset is proposed and validated over a novel set of breast cancer histone markers. At the beginning of this framework, different clustering algorithms are applied, and through a consensus clustering a set of common classes is defined in order to determine the fundamental characteristics of data expressed by different groups. Then these core groups may be assessed using supervised classification methods and characterised by the application of a set of visualisation techniques.

The paper is organised as follows: in Section 2, the proposed framework is presented and explained in detail. Section 3 is reserved for the experiment settings and the validation of the framework over a novel set of breast cancer histone markers provided by the Division of Molecular and Cellular Sciences, Centre for Biomolecular Sciences, School of Pharmacy at the University of Nottingham. In Section 4 a discussion of the results is reported together with directions for future research.

II. Strategy

The proposed framework needs several input sets of methods and parameters, and it is formed by different logical

Daniele Soria is with the School of Computer Sciences, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK (phone: +44 115 9514229; email: dqs@cs.nott.ac.uk).

Dr. Jonathan M. Garibaldi is with the School of Computer Sciences, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK (phone: +44 115 9514216; email: jmg@cs.nott.ac.uk).

steps which will be described below. In its most general parameterisation, the framework F may be written as

$$F(\Omega,\ P,\ C,\ V,\ K,\ B,\ S,\ a),$$

where the input arguments are as follows:

- The dataset under investigation Ω.
- The set of preliminary data analysis techniques and preprocessing algorithms P.
- The collection of several clustering techniques C which may be applied.
- The collection V of several validity indices which may be used to assess the grouping returned by cluster analysis.
- The set K of concordance measures (like the kappa coefficient of agreement, or Rand indexes).
- The collection B of visualisation techniques to characterise the groupings.
- The set of several supervised learning techniques S.
- The statistical coefficient a to assess the association between groups and variables of interest.

An organisation chart showing the overall approach and the logical steps used in this proposed pipeline is reported in Figure 1. Following this structure, each step of the framework is now presented.

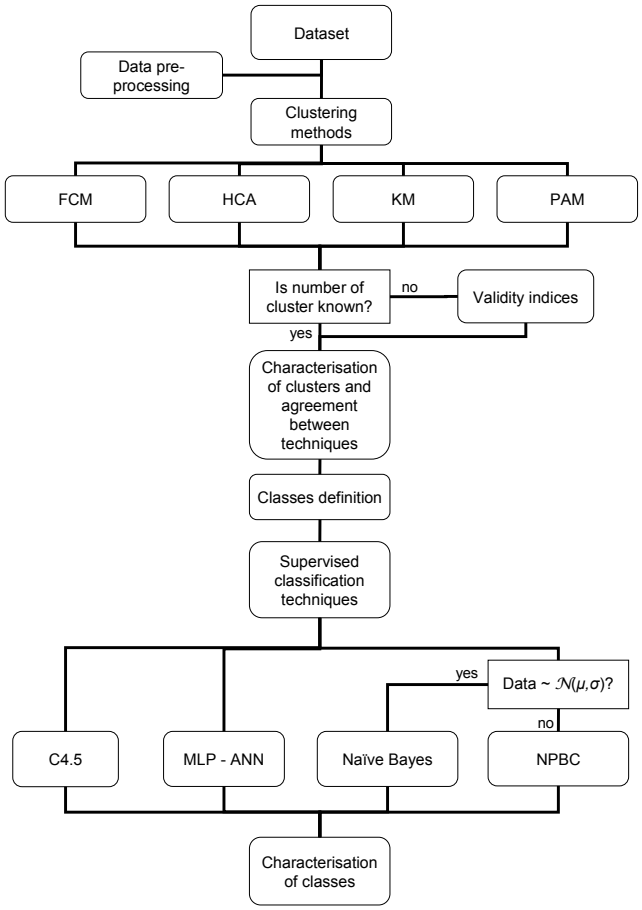


Fig. 1. Organisation chart of the proposed framework

1) In the first step, data preprocessing is performed. Rows which contain entries with missing values have to be deleted in order to run the clustering algorithms, and variables need to be 'homogeneous', which means that it is not convenient to have both numerical and categorical entries as part of the same variable distribution. If this happens, then clustering techniques may not be able to emphasise other possible structures within the dataset. In this 'data preprocessing' step, several descriptive statistics (minimum, maximum, mean, median, quartiles, etc.) need to be checked as well, in order to have a complete picture of the data under investigation and to immediately spot any inconsistency within them.
2) The second step involves clustering the data using a variety of algorithms over a range of numbers of clusters. Various unsupervised classification algorithms may be applied. In this framework, a subset of techniques from the set of four – hierarchical (HCA), K-means (KM), Fuzzy C-means (FCM) and Partitioning Around Medoids (PAM) – is selected to categorise cases into groups. Given that K-means and Fuzzy C-means methods are sensitive to cluster initialisation and in order to obtain reproducible results, these techniques are initialised with the cluster assignments obtained by hierarchical clustering. All of the above techniques have been previously analysed and used (see, for example [6], [5], [10]). This does not mean that these four are the best techniques to use, but they are among the most widely used clustering methods in machine learning and data mining as they performed quite well on a considerable amount of problems [5], [10], [11].
3) In this step, validity indices are applied to clustering results. One of the main problems related to cluster analysis is the choice of the number of clusters. To address this issue some external validation criteria (validity indices) may be used to compare one cluster solution to other cluster solutions and to choose the one suggested as optimal. Although there are many variations of validity indices, they are all either based on considering the data dispersion in a cluster and between clusters, or considering the scatter matrix of the data points and the one of the clusters centres. Several validity indices have been proposed in literature, but in this framework only few have been considered. The indices of Calinski and Harabasz [12], Hartigan [13], Scott and Symons [14], Marriot [15] and the two proposed by Friedman and Rubin [16] are used for partitional clustering (K-means and PAM). Instead, for the FCM method, the indices of Gath and Geva [17], Xie-Beni [18], and Bezdek [19] have been used. According to specific rules (see [20] and [21] for details), they all indicate the appropriate number of groups to consider in the analysis. When indices indicate different numbers, it is possible to use them to rank in order the suggested groupings and then take the minimum sum of ranks as a form of agreement between indices.

4) When clusters are returned, a general characterisation of them can be obtained through visualisation techniques. Biplots, which are built considering the first two principal components and represent clusters projected on them, are a useful tool as they provide a picture where the clusters have been 'spread out' as much as possible. Another technique for visualisation is the boxplot. It shows the distribution of each variable, computing its median value, the lower and upper quartiles and any outlier [22]. Through the computation of the boxplots of all variables divided by clusters and using the biplots as well, it is possible to obtain a first 'informal' description of the groupings obtained by the clustering techniques. In addition, the agreement between classifications returned is, in this guideline, assessed either using the Cohen's kappa and weighted kappa indices [23], [24], or the Rand and adjusted Rand indices [25], [26]. All these indices also give an indication about how likely will be to get a good consensus between classifications. Cohen's kappa index (κ) is a statistical measure of inter-rater agreement for qualitative (categorical) items [23]. It is generally thought to be a more robust measure than the percentage or proportion of agreement, since κ takes into account the agreement occurring by chance. Cohen's kappa is defined as

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

where p_o is the observed proportion of agreement, and p_c is the proportion of agreement expected by chance. Kappa takes negative values when there is less observed agreement than is expected by chance, zero when observed agreement can be (exactly) accounted for by chance, and one when there is complete agreement [23]. Cohen also introduced the *weighted kappa* index κ_w, considering the proportion of weighted disagreement. To find the latter, disagreement weight, v_{ij}, are defined by means of any judgment procedure set up to yield a ratio scale. It is convenient (even though not necessary) to assign zero to the 'perfect' agreement and the length of the vector of weights must equal the number of rating categories. Weighted kappa is then given by

$$\kappa_w = 1 - \frac{\sum v_{ij} p_{oij}}{\sum v_{ij} p_{cij}}.$$

where p_{oij} is the proportion of the joint judgments (N in number) observed in the ij cell, and p_{cij} is the proportion in the cell expected by chance. Like the 'unweighted' kappa index, κ_w is fully chance corrected. In this study, weights are set in decreasing order from one (perfect agreement) to zero (complete disagreement) and all levels disagreement between raters are weighted according to their distance from perfect agreement [24]. Another widely used measure to assess the agreement between classifications is the Rand index [25]. Given a set of objects S, suppose U and V represent two different partitions of the objects in S. Let a be the number of pairs of objects that are placed in the same element in partition U and in the same element in partition V, and d be the number of pairs of objects in different elements in partitions U and V [27]. The Rand index [25] is defined simply as the fraction of agreement, i.e.

$$R(U,V) = (a+d)/\binom{n}{2}.$$

The Rand index lies between 0 and 1, as, by definition, it is normalised. When the two partitions are identical, the Rand index is 1 [27].
However, the expected value of the Rand index of two random partitions does not take a constant value. For this reason, Hubert and Arabie [26] defined the *adjusted Rand index* which corrects for this by assuming the general form

$$\frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}.$$

In this general form the index is bounded above by 1, and takes the value 0 when the index equals its expected value [28]. As for the Rand index, a higher adjusted Rand index means a higher correspondence between the two partitions.

5) As per Figure 1, the next step is related to classes definition. This is done via a consensus clustering which may be performed in several ways. In this proposed framework, the classifications obtained by different clustering algorithms are used and, looking at the biplots, the cluster labels are aligned in order to have the same patient assigned to the cluster named in the same way by different algorithms. Looking then at the same cluster number / label across all methods, core classes are defined by taking into consideration those cases assigned to the same group by different methods. Two principles were used to guide the definition of consensus classes: (i) to include as many instances as possible and (ii) to take into account as many clustering algorithms as possible among the ones applied. However, it may happen [6] that these principles conflict, especially when the agreement between clustering methods is not high, and that the strict application of the second principle leads to a decrease in the number of patients assigned to classes. If this happens, it is then possible to employ a heuristic trade-off between the two principles [6].
6) To assess and verify the classes defined by the consensus clustering, supervised classification techniques may be used. Among them, the C4.5 classifier (C4.5), the MultiLayer Perceptron Artificial Neural Network (MLP-ANN) and the naive Bayes classifier are considered in this framework. When data do not follow a normal distribution, a 'non-parametric' Bayesian classifier (NPBC) (recently developed and presented in [29]) may be used.
7) In the last step, the identified core classes are described resorting again to biplots and boxplots. When computing the biplots of classes, the 'not classified' cases

usually are concentrated in the middle of the region. In addition, the correlation between classes and particular features of interest is computed resorting to the Phi (ϕ) statistics [30].

III. Validation

A. Experiment settings

To validate the approach presented in the previous section, the framework was applied in the following configuration: (Ω_1, P_1, C_1, V_1, K_1, B_1, S_1, ϕ) where each input set is now described.

- Ω_1 = A novel dataset provided by the School of Pharmacy at the University of Nottingham.
- $P_1 = \{$Deletion of rows where missing values appear, descriptive statistics computation$\}$.
- $C_1 = \{$KM, PAM, FCM$\}$.
- $V_1 = \{$The same validity indices reported in the previous Section and already used in [6] and [21]$\}$.
- $K_1 = \{\kappa, \kappa_w\}$.
- $B_1 = \{$Biplots, boxplots$\}$.
- $S_1 = \{$C4.5$\}$.
- ϕ as the index to assess the association between classes and clinical variables available.

The choice of using the above configuration is motivated by the robustness of the clustering algorithms and by the C4.5 producing a set of rules easily understandable by clinicians, which are usually not familiar with computational analysis.

B. Case study

The dataset Ω_1 used to validate the proposed approach was a collection of 1254 consecutive breast tumours diagnosed from 1986 to 1998 included in the Nottingham Tenovus Primary Breast Carcinoma Series. Full details of the characterisation of the tissue microarray and the cohort of the patients are described in [31], [32]. Survival data were maintained on a prospective basis. Breast cancer specific survival was taken as the time (in months) from the date of the primary surgical treatment to the time of death from breast cancer [32]. A grading score was also available in this dataset. Grade is one of the components of the Nottingham Prognostic Index [33] and is determined by the microscopic evaluation of tumour cells by pathologists [34], [35].

Breast cancer tissue microarrays were prepared and immunohistochemically stained to detect four histone markers as described in [36]. Each case was sampled twice from both the centre and the periphery of the tumour. The histone markers selected for this study were hMOF, ACH4K16, H3K9Me3 and SUV. They all have different functions: hMOF is a histone transferase enzyme which is responsible for H4K16 acetylation. ACH4K16 is a marker of active gene, while H3K9Me3 is a marker of silenced gene. Finally, SUV is the main factor responsible for H3K9 tri-methylation [37].

C. Results

This collection of data presented many missing values; for the analysis described below, the four histone markers were only considered as well as those patients for which all the information was present, thus reducing the number of patients to 347. The basic descriptive statistics like minimum, mean and maximum values for each feature were computed and together with the deletion of all rows where missing values were found, they formed the pre-processing techniques of the P_1 input set.

To assess the grouping, the KM, PAM and FCM algorithms (see [38], [39], [40]) were applied with the number of clusters varying between two and twenty (the number of clusters is an explicit input parameter for all algorithms).

For the partitional clustering (KM and PAM), the same validity indices used in [6] were used for these experiments, as well as the decision rules reported in Table 1 of the same paper. The values of the indices for both K-means and PAM, for 2 to 20 clusters are shown in Figure 2; (a) shows the validity decision rule values obtained for K-means and (b) shows those obtained for PAM. The best number of clusters according to each validity index, for each clustering algorithm, is shown in Table I. This corresponds to either the maximum or the minimum decision rule value (depending on the index), as indicated by the solid circle in Figure 2.

TABLE I
OPTIMUM NUMBER OF CLUSTERS ESTIMATED BY EACH INDEX FOR K-MEANS AND PAM METHODS

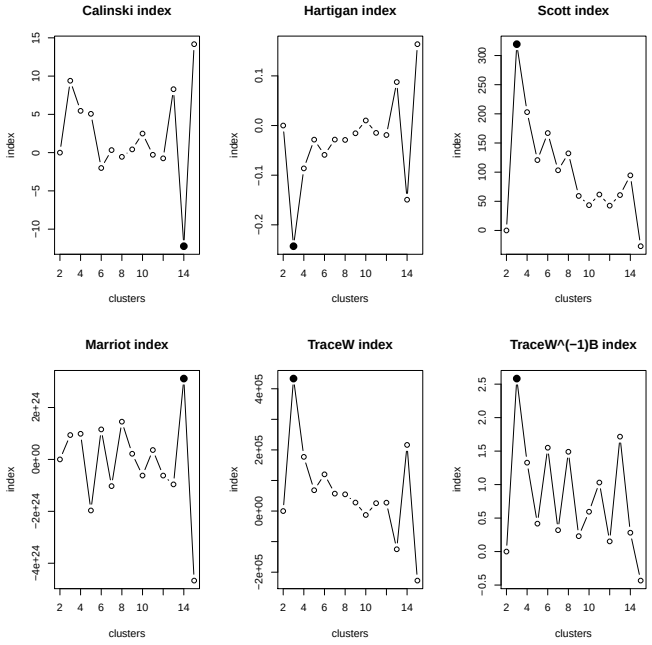
Index	K-means	PAM
Calinski and Harabasz	14	3
Hartigan	3	3
Scott and Symons	3	3
Marriot	14	3
TraceW	3	3
Trace$W^{-1}B$	3	3
Minimum sum of ranks	3	3

When FCM was applied, four validity indices were considered, and their values for 2 to 20 clusters are shown in Figure 3. The best number of clusters according to each validity index, for FCM clustering algorithm, is shown in Table II. This corresponds to either the index maximum or minimum, as indicated by the solid circle in Figure 3.

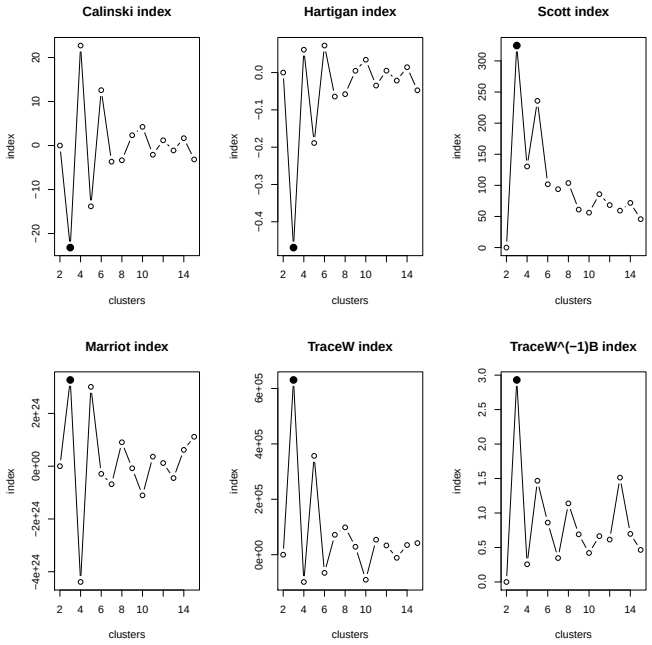
TABLE II
OPTIMUM NUMBER OF CLUSTERS ESTIMATED BY EACH INDEX FOR FCM METHOD

Index	Fuzzy C-means
Fuzzy Hypervolume	2
Partition Density	3
Xie-Beni	2
Partition Coefficient	2
Minimum sum of ranks	2

From Table I it can be seen that all indices applied to the PAM results suggested three groups, while such an agreement was not evident in the case of K-means algorithm. However, resorting to the minimum sum of ranks for the indices, it could be observed that both methods indicated three as the



(a) K-means indices behaviors



(b) PAM indices behaviors

Fig. 2. Cluster validity indices obtained for K-means and PAM clustering, for varying cluster numbers from 2 to 20

best number of clusters. For the Fuzzy C-means algorithm, instead, the minimum sum of ranks for the indices indicated two clusters. However, as three was the second best number of clusters and in order to be consisted with KM and PAM results, three groups were also considered for the Fuzzy C-means algorithm. The cluster distribution (number of patients in each cluster) obtained for the K-means, PAM and FCM methods is reported in Table III.

The correspondence of patients assigned in the three

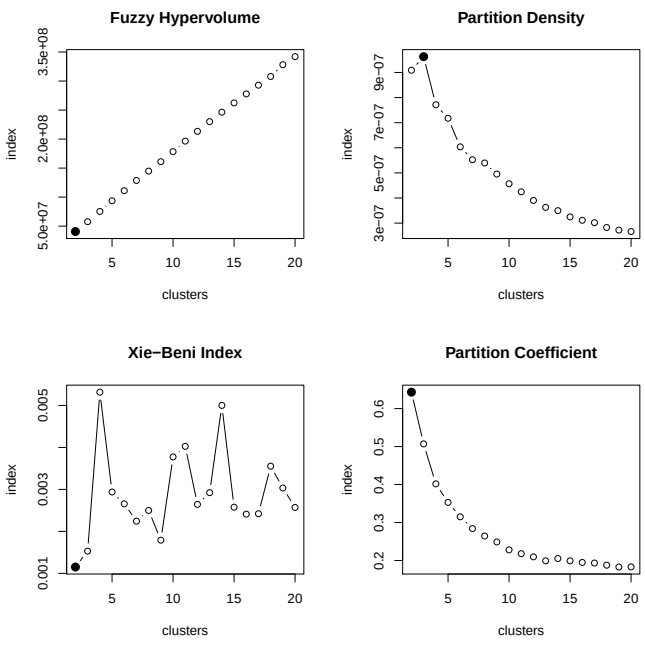


Fig. 3. Cluster validity indices obtained for Fuzzy C-means clustering, for varying cluster numbers from 2 to 20

TABLE III

NUMBER OF CASES IN EACH CLUSTER

Cluster	K-means	PAM	FCM
1	144	161	136
2	105	96	109
3	98	90	102

TABLE IV

KAPPA AND *weighted kappa* INDEX FOR DIFFERENT CLASSIFICATIONS

	K-means	PAM
FCM	0.926	0.89
	0.91	*0.88*
K-means	—	0.911
	—	*0.906*

clusters solution for each of the methods was then examined resorting to both the unweighted and weighted kappa index κ. For the weighted-kappa index, weights were set in decreasing order from one (perfect agreement) to zero (complete disagreement) with a 0.5 step between levels. Results are reported in Table IV. From this table, an almost perfect agreement between the three techniques is visible.

Focusing on the cluster correspondences, core classes containing the biggest possible number of patients were defined. Considering the agreement among the clustering techniques and looking at those patients assigned to the same group by the different methods, three common classes were created containing the 91.1% of the overall population. In practice, 31 patients were not assigned to any of these three classes and were placed into a 'not classified' (NC) group. The distribution of patients in the three 'common' classes is reported in Table V, together with the rule applied to define each class.

TABLE V

DISTRIBUTION OF PATIENTS IN THE 'COMMON' CLASSES

Class	No. of cases
1 (KM1 ∧ PAM1 ∧ FCM1)	132
2 (KM2 ∧ PAM2 ∧ FCM2)	95
3 (KM3 ∧ PAM3 ∧ FCM3)	89
Total number of cases assigned to classes 1 – 3	316
Total number of cases not classified	31

Biplots of the three consensus classes were produced and are reported in Figures 4 and 5, which provide a visualisation of the classes projected on the first two principal components. The arrows in the plots represent the variables (markers) and their direction indicate in which group they are more expressed.

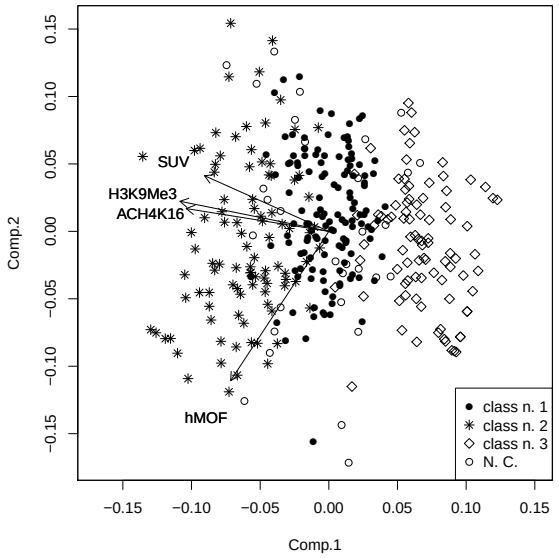


Fig. 4. Biplots of classes projected on the first and second principal component axes for all patients

Figure 4 shows the biplot obtained for all patients, in which the cases not assigned to any class (NC) are represented by empty circles. It can be seen that these fall mainly into the centre region of the biplot. Figure 5 shows the biplot obtained for only those patients assigned to classes 1 – 3. The first axis was mainly determined, on the left, by ACH4K16 and H3K9Me3 markers, while the second one is determined, on the bottom, by hMOF over-expression.

Figure 6 shows boxplots of all four markers, (a) for those cases assigned to classes 1 to 3, and (b-d) for each class separately.

By visual inspection of both the biplots and the boxplots, an 'informal' description of each class could be derived. It seems quite evident that class 3 is mainly characterised by low expression of all the four markers. Compared to the overall distribution, class 2 appears to express higher values while class 1 is quite similar, especially with respect to hMOF and ACH4K16.

Starting from these consensus/common data, we investigated whether it was possible to establish a set of rules

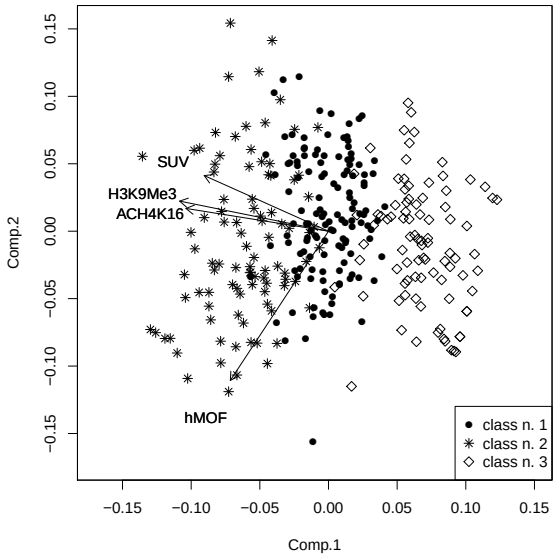
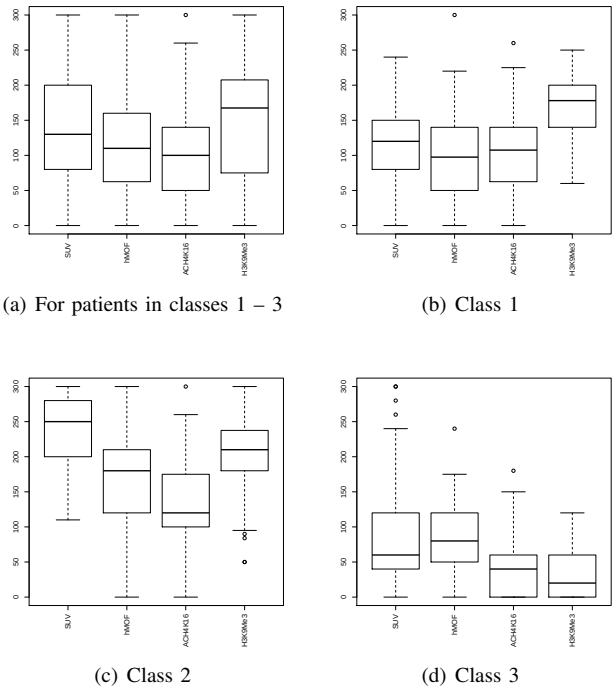


Fig. 5. Biplots of classes projected on the first and second principal component axes for patients in classes 1 – 3



(a) For patients in classes 1 – 3 (b) Class 1

(c) Class 2 (d) Class 3

Fig. 6. Boxplot for all markers grouped by class

to determine in which group a patient is more likely to be assigned starting from its variables values. To do so, supervised machine learning algorithms were used (results not shown).

As mentioned before, several items of clinical information were also available for this study. In particular, the overall survival of patients was considered, and using the Kaplan-Meier estimator [41], [42] a curve of the predicted survival against time for each class was produced. The Kaplan-Meier curves obtained for this study are reported in Figure 7. It is important to note that several rows, representing patients

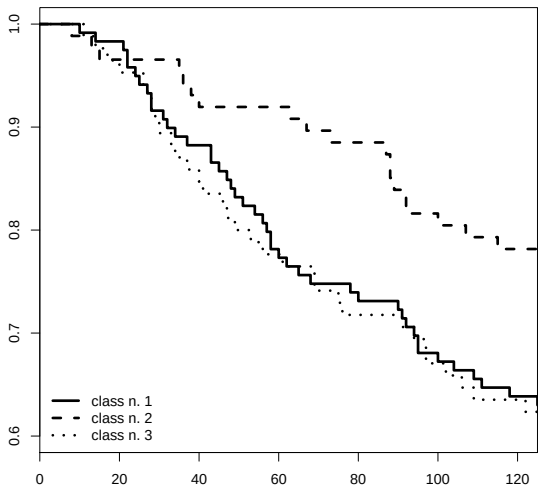


Fig. 7. Kaplan - Meier curves for months of survival divided by class

with missing information about survival time and recurrence, were deleted for computation of the curves, leaving the total number of patients equal to 291. The best overall survival is visible for patients grouped in class 2, which is characterised by high values of covariates. Classes 1 and 3 have the worst, equally poor, survival. These results are in agreement with those previously obtained in [32] and [43] where high values of hMOF and ACH4K16 were associated with a better survival.

As a last analysis, the association between tumour grade [34], [35] and classes was assessed, resorting to the Phi (ϕ) statistics [30]. This association is reported in Table VI where the total number of patients is 313, as there were 3 missing information for grade.

TABLE VI
COMMON CLASSES DISTRIBUTION IN RELATION TO GRADING SCORE

	Common classes			
	1	2	3	ϕ
Low Grade	30	31	11	
Interm. Grade	33	42	30	0.293
High Grade	67	22	47	

Although the Phi statistic is low, an interesting result may still be inferred from Table VI. As a matter of fact, it can be seen that the majority of high grade patients, which are known to have a poor prognosis [34], are grouped in classes 1 and 3, which, according to the Kaplan-Meier curves reported in Figure 7 are the groups with the worst overall survival. This proves that the common classes, which are derived only on the basis of the four markers and without using any clinical data, are able to group together patients with similar outcome.

IV. DISCUSSION

In this paper we have presented a novel framework to allow the elucidation of core classes within a dataset. It follows a logical scheme in which at the beginning unsupervised clustering techniques are used to group patients (or any kind of data) in clusters which share similar characteristics. It is important to use more than a single clustering method, as it has been proved in literature [5], [6] that different algorithms return different groups and it is not possible to say which, if it really exists, is the best clustering technique to use. By a visual inspection of the results and by using an index to assess the degree of agreement between different classifications, an 'informal' consensus clustering may be derived, considering those patients assigned to the same group by different algorithms as 'in-class' and labelling all the others as 'not classified'. The resulting classes may be then analysed in different ways, either using biplots and boxplots, or looking at their relations with other variables (which, in this study, were several clinical information). In any of those cases, automated supervised classification techniques may be used to confirm and assess the identified grouping.

We validated the proposed approach on a dataset of breast cancer histone biomarkers which was provided by the Division of Molecular and Cellular Sciences in the School of Pharmacy at the University of Nottingham. The three common classes identified by the consensus clustering have a quite clear definition. In fact the three groups are somehow characterised by low / intermediate / high markers levels. Moreover, the agreement between classifications (kappa and weighted kappa indexes) is very high. From a clinical point of view, the second class presents a higher survival rate than the other two, so leading to the conclusion that the higher the values of biomarkers, the better the prognosis for the patient. A similar result was reported in [32], where more histone markers were considered and just two clustering algorithms were applied.

Although the results obtained so far still need afinal interpretation from clinicians and researchers at the Schools of Pharmacy, this study served to present and validate our proposed procedure, which, so far, has given very promising and encouraging results.

As a future work different data sets will be used to validate again the framework. In addition, model-based clustering approaches and semi-supervised learning techniques will be considered and their inclusion in the proposed framework will be evaluated. Finally, this work will be extended to include comparisons to other approaches recently developed (e.g. [44]).

ACKNOWLEDGMENTS

The authors would like to thank Prof. David M. Heery and his Ph.D. student Magdy Korashy Abdel Fatah for providing the data used in this work. This study was, in part, supported by the BIOPTRAIN FP6 Marie-Curie EST Fellowship (FP6-007597).

REFERENCES

[1] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, pp. 91–118, 2003.

[2] C. Perou, T. Sørlie, M. Eisen, M. Van De Rijn, S. Jeffrey, C. Rees, J. Pollack, D. Ross, H. Johnsen, L. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S. Zhu, P. Lonning, A. Børresen-Dale, P. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747–752, 2000.

[3] T. Sørlie, C. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. Eisen, M. Van De Rijn, S. Jeffrey, T. Thorsen, H. Quist, J. Matese, P. Brown, D. Botstein, P. Eystein Lonning, and A. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc Natl Acad Sci U S A*, vol. 98, pp. 10 869–10 874, 2001.

[4] L. Van't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2002.

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

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

[7] P. Kellam, X. Liu, N. Martin, C. Orengo, S. Swift, and A. Tucker, "Comparing, contrasting and combining clusters in viral gene expression data," in *Proceedings of 6th Workshop on Intelligent Data Analysis in Medicine*, 2001.

[8] V. Filkov and S. Skiena, "Integrating microarray data by consensus clustering," in *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp. 418– 426.

[9] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam, "Consensus clustering and functional interpretation of gene-expression data," *Genome Biology*, vol. 5:R94, 2004.

[10] X. Wang and J. Garibaldi, "A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis," in *Proceedings of second international conference in Computational Intelligence in Medicine and Healthcare*, 2005, pp. 250–256.

[11] R. Diallo-Danebrock, E. Ting, O. Gluz, A. Herr, S. Mohrmann, H. Geddert, A. Rody, K. Schaefer, S. Baldus, A. Hartmann, P. Wild, M. Burson, H. Gabbert, U. Nitz, and C. Poremba, "Protein expression profiling in high-risk breast cancer patients treated with high-dose or conventional dose-dense chemotherapy," *Clin Cancer Res*, vol. 13, pp. 488–497, 2007.

[12] R. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communs statist*, vol. 3, pp. 1–27, 1974.

[13] J. Hartigan, *Clustering Algorithms*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1975.

[14] A. Scott and M. Symons, "Clustering methods based on likelihood ratio criteria," *Biometrics*, vol. 27, no. 2, pp. 387–397, 1971.

[15] F. Marriot, "Practical problems in a method of cluster analysis," *Biometrics*, vol. 27, no. 3, pp. 501–514, 1971.

[16] H. Friedman and J. Rubin, "On some invariant criteria for grouping data," *Journal of the American Statistical Association*, vol. 62, no. 320, pp. 1159–1178, 1967.

[17] I. Gath and A. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773–781, 1989.

[18] L. Xie and G. Beni, "Validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[19] J. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, pp. 191–203, 1984.

[20] A. Weingessel, E. Dimitriadou, and S. Dolnicar, "An examination of indexes for determining the number of clusters in binary data sets," Working Paper No.29, 1999.

[21] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, pp. 107–145, 2001.

[22] P. Velleman and D. Hoaglin, *Applications, Basics and Computing of Exploratory Data Analysis*. Boston, Mass.: Duxbury Press, 1981.

[23] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[24] ——, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, pp. 213–220, 1968.

[25] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.

[26] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[27] K. Yeung and W. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[28] ——, "An empirical study on principal component analysis for clustering gene expression data," Department of Computer Science & Engineering, University of Washington, Seattle, US, Tech. Rep., 2000.

[29] D. Soria, J. Garibaldi, F. Ambrogi, E. Biganzoli, and I. Ellis, "A 'non-parametric' version of the naive bayes classifier," *Submitted to Data & Knowledge Engineering*, 2010.

[30] B. Everitt, *The Cambridge Dictionary of Statistics*. Cambridge University Press, 2002.

[31] S. Elsheikh, A. Green, M. Lambros, N. Turner, M. Grainge, D. Powe, I. Ellis, and J. Reis-Filho, "FGFR1 amplification in breast carcinomas: A chromogenic in situ hybridisation analysis," *Breast Cancer Research*, vol. 9:R23, 2007.

[32] S. Elsheikh, A. Green, E. Rakha, D. Powe, R. Ahmed, H. Collins, D. Soria, J. Garibaldi, C. Paish, A. Ammar, M. Grainge, G. Ball, M. Abdelghany, L. Martinez-Pomares, D. Heery, and I. Ellis, "Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome," *Cancer Research*, vol. 69, pp. 3802–3809, 2009.

[33] M. Galea, R. Blamey, C. Elston, and I. Ellis, "The Nottingham Prognostic Index in primary breast cancer," *Breast Cancer Res Treat*, vol. 22, pp. 207–219, 1992.

[34] I. Ellis, M. Galea, N. Broughton, A. Locker, R. Blamey, and C. Elston, "Pathological prognostic factors in breast cancer. II. histological type. Relationship with survival in a large study with long-term follow-up," *Histopathology*, vol. 20, pp. 479–489, 1992.

[35] E. Rakha, M. El-Sayed, A. Lee, C. Elston, M. Grainge, Z. Hodi, R. Blamey, and I. Ellis, "Prognostic significance of nottingham histologic grade in invasive breast carcinoma," *J Clin Oncol*, vol. 26, no. 19, pp. 3153–3158, 2008.

[36] D. Abd El-Rehim, S. Pinder, C. Paish, J. Bell, R. Blamey, J. Robertson, R. Nicholson, and I. Ellis, "Expression of luminal and basal cytokeratins in human breast carcinoma," *Journal of Pathology*, vol. 203, pp. 661–671, 2004.

[37] D. Maglott, J. Ostell, K. Pruitt, and T. Tatusova, "Entrez Gene: Gene-centered information at NCBI," *Nucleic Acids Research*, vol. Database Issue, pp. D54–D58, 2005.

[38] J. MacQueen, "Some methods of classification and analysis of multivariate observations," in *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley*, 1967, pp. 281–297.

[39] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1990.

[40] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, plenum, New York ed., 1981.

[41] E. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[42] J. Kalbfleisch and R. Prentice, *The Statistical Analysis of Failure Time Data*, 2nd ed. Hoboken, N.J.: Wiley-Interscience, 2002.

[43] S. Pfister, S. Rea, M. Taipale, F. Mendrzyk, B. Straub, C. Ittrich, O. Thuerigen, H. Sinn, A. Akhtar, and P. Lichter, "The histone acetyltransferase hMOF is frequently downregulated in primary breast carcinoma and medulloblastoma and constitutes a biomarker for clinical outcome in medulloblastoma," *Int. J. Cancer*, vol. 122, no. 6, pp. 1207–1213, 2008.

[44] A. Fred and A. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, 2005.