Soria, Daniele and Garibaldi, Jonathan M. and Ambrogi, Federico and Lisboa, Paulo J.G. and Boracchi, Patrizia and Biganzoli, Elia M. (2008) Clustering breast cancer data by consensus of different validity indices. In: International Conference on Advances in Medical, Signal and Information Processing (4th), 14-16 July 2008, Santa Margherita Ligure, Italy.

CLUSTERING BREAST CANCER DATA BY CONSENSUS OF DIFFERENT VALIDITY INDICES

D. Soria[*], **J.M. Garibaldi**[*], **F. Ambrogi**[†], **P.J.G. Lisboa**[♯], **P. Boracchi**[†], **E. Biganzoli**[†]

[*]School of Computer Science, University of Nottingham, UK
[†]Italian National Cancer Institute, University of Milan, Italy
[♯] School of Computing and Mathematical Sciences, Liverpool John Moores University, UK

Keywords: Clustering algorithms, Breast cancer, Validity indices.

Abstract

Clustering algorithms will, in general, either partition a given data set into a pre-specified number of clusters or will produce a hierarchy of clusters. In this paper we analyse several different clustering techniques and apply them to a particular data set of breast cancer data. When we do not know a priori which is the best number of groups, we use a range of different validity indices to test the quality of clustering results and to determine the best number of clusters. While for the K-means method there is not absolute agreement among the indices as to which is the best number of clusters, for the PAM algorithm all the indices indicate 4 as the best cluster number.

1 Introduction

Clustering is the process of dividing data elements into classes (clusters) so that items in the same class are as similar as possible. Since there is no *a priori* fixed method to determine the best number of clusters for any given data set, a number of cluster validity indices have been proposed which attempt to measure how 'good' a particular clustering assignment is. Conceptually, they measure the 'compactness' of each cluster and the 'separation' of cluster centres. A solution in which all data are assigned to clusters such that they are close to the cluster centre while all the clusters are far apart may be considered a good solution.

Numerous studies have reported distinct breast cancer groups based on gene expression profiles. Recently we have instead investigated an alternative approach based on immunocytochemistry, an established robust technology, in a total of 1,076 invasive breast cancer cases [1]. Using a hierarchical clustering methodology and Artificial Neural Network (ANN) modelling techniques, six groups with distinct patterns of protein expression were identified.

In this paper, we present a further study in which we extended our previous work by examining a range of alternative clustering techniques and utilising cluster validity indices to externally verify the number of cluster groups arrived at.

2 Patients and Methods

2.1 Patients

A series of 1076 patients from the Nottingham Tenovus Primary Breast Carcinoma Series presenting with primary operable invasive breast cancer between 1986-98 were used. For clustering analyses, we used a large panel of tumour markers, which are listed in [1]. Most of the proteins selected to study in our work have a well-established role in breast carcinogenesis.

2.2 Methods

Five different algorithms were used for the cluster analysis: (i) Hierarchical (H), (ii) Fuzzy C-Means (FCM), (iii) K-means initialized with hierarchical clustering (method average), (iv) Partitioning Around Medoids (PAM), and (v) Adaptive Resonance Theory (ART).

2.2.1 Hierarchical

Hierarchical clustering builds (agglomerative), or breaks up (divisive), a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements at one end and a single cluster containing every element at the other. Cutting the tree at a given height will give a clustering at a selected precision.

The hierarchical method was used as described previously in [1] and the same (six) cluster assignments as obtained were used.

2.2.2 Fuzzy C-Means

The FCM algorithm is based on the minimization of an objective function $J(U, V)$ (1) to achieve a good classification.

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{i,j})^m \|x_i - v_j\|^2 \qquad (1)$$

In (1) the expression $X = \{x_1, x_2, ..., x_n\}$ is a collection of data, where n is the number of data points and $V = \{v_1, v_2, ..., v_c\}$ is the set of corresponding cluster centres in the data set X, where c is the number of clusters. μ_{ij} is the membership degree of data x_i to the cluster centre v_j ($\mu_{i,j} \in [0, 1]$). m is called the "fuzziness index" and the value of $m = 2.0$ is usually chosen. A full description of this method can be found in [3].

2.2.3 K-means

The K-means technique aims to partition the data into k groups such that the sum of squares from points to the assigned cluster centres is minimized. As for the fuzzy c-means, an objective function $J(V)$ (2) should be minimized, but for this method it has the following aspect:

$$J(V) = \sum_{j=1}^{k} \sum_{i=1}^{c_j} ||x_i - v_j||^2 \qquad (2)$$

where $||x_i - v_j||$ is the Euclidean distance between x_i and v_j and c_j is the number of data points in the cluster j. The j-th centre v_j can be calculated as:

$$v_j = \frac{1}{c_j} \sum_{i=1}^{c_j} x_i, \qquad j = 1, ..., k. \qquad (3)$$

K-means clustering is dependent on the initial setting of the cluster assignments (which, in turn, determines the initial cluster centres). Various techniques have been proposed for the initialisation of clusters [2], but for this study we used a fixed initialisation of the cluster assignment obtained with hierarchical clustering.

2.2.4 Partitioning Around Medoids

The PAM algorithm is based on the search for k representative objects (the so-called *medoids*) among the observations of the data set. These observations should represent the structure of the data. After finding a set of k medoids, k clusters are constructed by assigning each observation to the nearest medoid. The goal is to find k representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object. Dissimilarities are non-negative numbers that are close to zero when two points are near to each other and that become large when they are very different.

2.2.5 Adaptive resonance theory

The adaptive resonance theory (ART) algorithm for self-organisation [4] was motivated by analogy with biological nervous systems, where the array of memory prototypes is thought to grow in a stable manner in the presence of new information, without necessarily over-writing previously derived states. The ART algorithm has two characteristic properties; firstly, it constrains the self-organised groups by a maximum separation from any point to the group prototype, creating new prototypes dynamically as necessary and, secondly, it introduces a bias pushing the group prototype toward the covariate axes, which tends to increase the differentiation between clusters. For details of the algorithm, see [4].

2.2.6 Validity Indices

Clustering validity is a concept that is used to evaluate the quality of clustering results. If the number of clusters is not known prior to commencing an algorithm, a clustering validity index may be used to find the optimal number of clusters for a given data set. Although there are many variations of validity indices, they are all based on considering the data dispersion in a cluster and between clusters.

For hierarchical clustering, the same six clusters as obtained previously ([1]) were utilised without further examining cluster validity.

For the Fuzzy C-Means algorithm Gath-Geva ([6]), Xie-Beni ([11]), Partition Coefficient and Partition Entropy ([3]) indices have been used.

For K-means and PAM clustering, the algorithms were both run for between 2 and 20 clusters. After each iteration, six cluster validity indices specified below were calculated and recorded, in order to determine the best number of clusters. The indices are ([10]): Calinski and Harabasz, Hartigan, Scott and Symons, Marriot, TraceW, and TraceW^{-1}B. For each index the number of clusters to be considered was chosen according to the rule reported in [10].

For the remaining technique (ART), the cluster number is a fixed parameter of the algorithm — i.e. the cluster number is provided to the algorithm, which then attempts to find the best assignment of the data to the given number of clusters (while determining the location of each cluster centre).

2.2.7 Visualisation

To enable visualisation, the original data space (the 25-dimensional space of protein markers) was transformed by principal component analysis (PCA) [7], and then the points were plotted at their projected position against the first and second principal components' axes. Such a plot provides a picture in which the clusters have been 'spread out' as much as possible according to the first two components. We also used different colors for patients belonging to different clusters.

All our work was done using R, which is a free software environment for statistical computing and graphics. [9]

3 Results

3.1 Clustering Results

3.1.1 Hierarchical and ART results

As the Hierarchical method was exactly as we used previously ([1]), the same six clusters were obtained.

For the ART algorithm, the parameters of the model were adjusted to result in the same number of clusters as the other methods, so facilitating the identification of concordant cluster membership across the different approaches.

3.1.2 Fuzzy C-means results

The results for this method indicated that fuzzy c-means algorithm was not obtaining good cluster partitions and, instead, was assigning all data points to all clusters with equal membership. Furthermore, it was found that when the data set was divided into more than three clusters, the final clustering obtained assigned no data to some clusters (i.e. some clusters had no elements). Additional validity indices ([10, 8]) were also calculated based on the final hard clusters, but without improvement in results. As a consequence, fuzzy c-means was dropped from further analysis.

3.1.3 K-means and PAM results

The validity indices were calculated for each method, for 2 to 20 clusters. The corresponding best number of clusters is shown in Table 1.

It can be seen that, while there was not absolute agreement among the indices as to which was the best number of clusters for the K-means method, there is good agreement that the best number of clusters for the PAM algorithm is four. But, on further inspection, it can be seen that even for the K-means, there is more agreement than might be immediately apparent. For example, the Scott and Symons index (which indicated that the best number of clusters was three) indicated that the second best number of clusters was six. Consequently, the indices were used to rank order the number of clusters and the minimum sum of ranks was examined. It was found that the minimum sum of ranks (a form of consensus among the indices) indicated that the overall best number of clusters was six for K-means.

	K-means	*PAM*
Calinski and Harabasz	6	4
Hartigan	3	4
Scott and Symons	3	4
Marriot	6	4
TraceW	4	4
Friedman and Rubin	3	4
Minimum sum of ranks	6	4

Table 1: Optimum number of clusters estimated by each index for K-means and PAM methods

A summary of the cluster distributions (number of patients in each cluster) obtained for each of the methods is shown in Table 2.

	PAM		*Hierarchical*	*K-means*	*ART*
1	382	1	336	282	238
2	324	2	180	301	408
3	153	3	234	134	188
4	217	4	4	138	96
		5	183	124	35
		6	139	97	111

Table 2: Number of cases in each cluster

3.2 Visualisation

Biplots of the clusters obtained for each method, as shown in Figure 1, were produced. From these plots, it can be seen that the first axis splits two clusters (1 & 2) over the left-hand side of the biplots. A third cluster (cluster 3) is evident towards the bottom of the biplots. The PAM method places all patients on the right-hand side into a single cluster (PAM cluster 4), while for the other methods, various splits of these data into three clusters (4, 5 & 6) can be seen.

4 Discussion

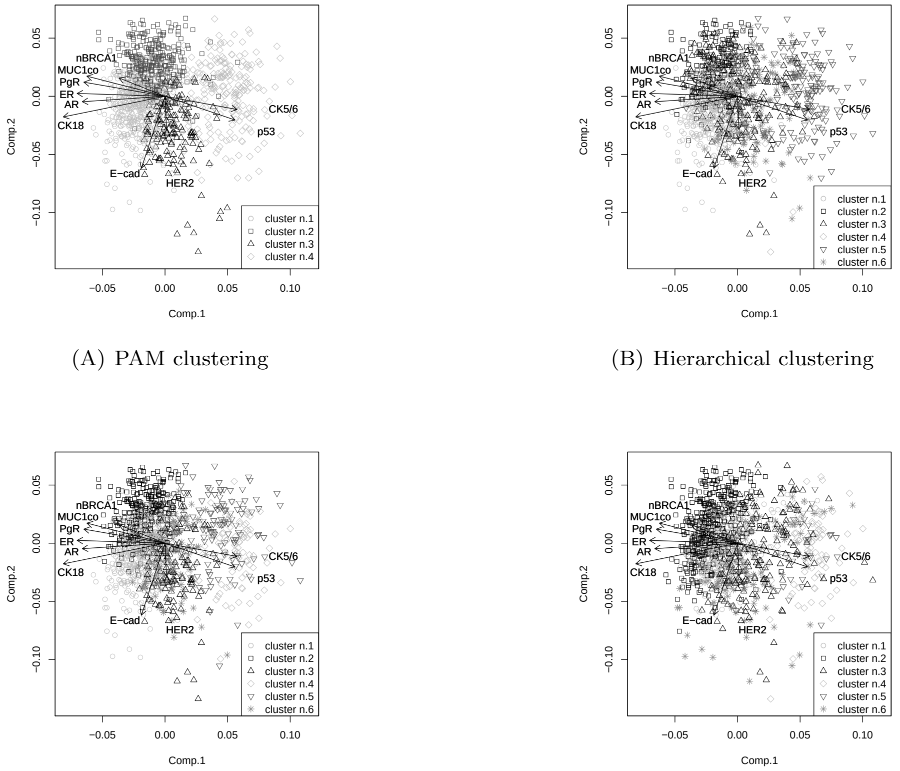
In this paper we reviewed five different clustering techniques and applied them to a particular case study in combination with a range of cluster validity indices. From our experiments we found different results for each of them.

The fuzzy c-means method seemed to be the weakest one among the proposed techniques. It did not return a clear classification and the membership function for each datum was very poor. Even the validity indices computation did not suggest any relevant result.

We also found a difference between the most similar methods: in fact, although K-means and PAM share several features, they ended up with different results. Looking at the validity indices values, we found that PAM algorithm suggested a clear classification in four groups, while the K-means one was more unstable. This difference might be explained saying that PAM is a more robust method compared to K-means, as it minimises a sum of dissimilarities (real numbers), instead of a sum of squared euclidean distances. On further inspection, using the minimum sum of ranks of validity indices, we found that we could choose six clusters for the K-means method. This fact was in accordance with previous results (hierarchical clustering) and with the last method applied (ART), which needed the number of clusters as an input value.

We then compared PAM method (with four clusters division) with the other ones (Hierarchical, K-Means and ART all with six clusters split) using the biplots (Fig. 1). We note that the four PAM clusters are more compact and better separated compared to other methods. We have also carried out a detailed examination of how these clusters relate to clinical factors [5].

In conclusion, using five different clustering methodologies in combination with a range of cluster validity indices, we have found that there are two possible ways of splitting our data, one using four groups and the other using six. It is worth noting that in such a large, complex, high-dimensional data set, it is extremely unlikely that a wide range of clustering algorithms would reach perfect agreement.



(A) PAM clustering

(B) Hierarchical clustering

(C) K-means clustering

(D) ART clustering

Figure 1: Biplots of clusters projected on the first and second principal component axes

Acknowledgements

This study was supported by the BIOPATTERN FP6 Network of Excellence (FP6-IST-508803) and the BIOPTRAIN FP6 Marie-Curie EST Fellowship (FP6-007597).

References

[1] D.M. Abd El-Rehim, G. Ball, S.E. Pinder, E. Rakha, C. Paish, J.F. Robertson, D. Macmillan, R.W. Blamey, and I.O. Ellis. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int. Journal of Cancer*, 116:340–350, 2005.

[2] M. Al-Daoud and S. Roberts. New methods for the initialisation of clusters. *Pattern Recognition Letters*, 17(5):451–455, 1996.

[3] J.C. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–73, 1974.

[4] G.A. Carpenter and S. Grossberg. ART2: Stable self-organization of pattern recognition codes for analogue input patterns. *Applied Optics*, 26:4919–4930, 1987.

[5] J.M. Garibaldi, D. Soria, F. Ambrogi, A.R. Green, D. Powe, E. Rakha, R.D. Macmillan, R.W. Blamey, G. Ball, P.J.G. Lisboa, T.A. Etchells, P. Boracchi, E. Biganzoli, and I.O. Ellis. Identification of key breast cancer phenotypes. *Submitted to International Journal of Cancer*, 2008.

[6] I. Gath and A.B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–781, 1989.

[7] J.E. Jackson. *A User's Guide to Principal Components*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1991.

[8] L. Kaufman and P.J. 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.

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

[10] 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.

[11] L.X. Xie and G. Beni. Validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.