Computers in Biology and Medicine 40 (2010) 318–330

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients

Daniele Soria [a,*], Jonathan M. Garibaldi [a], Federico Ambrogi [c], Andrew R. Green [b], Des Powe [b], Emad Rakha [b], R. Douglas Macmillan [f], Roger W. Blamey [f], Graham Ball [d], Paulo J.G. Lisboa [e], Terence A. Etchells [e], Patrizia Boracchi [c], Elia Biganzoli [c], Ian O. Ellis [b]

[a] School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK
[b] School of Molecular Medical Sciences, Nottingham University Hospitals and University of Nottingham, Queens Medical Centre, Derby Road, Nottingham NG7 2UH, UK
[c] Institute of Medical Statistics and Biometry, University of Milan, Via Venezian 1, 20133 Milan, Italy
[d] School of Science and Technology, Nottingham Trent University, Clifton Campus, Clifton Lane, Nottingham NG11 8NS, UK
[e] School of Computing and Mathematical Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK
[f] The Breast Institute, Nottingham City Hospital, Hucknall Road, Nottingham NG5 1PB, UK

ARTICLE INFO

Article history:
Received 18 June 2009
Accepted 4 January 2010

Keywords:
Breast cancer
Molecular classification
Clustering methods
Consensus clustering
Validity indices

ABSTRACT

Single clustering methods have often been used to elucidate clusters in high dimensional medical data, even though reliance on a single algorithm is known to be problematic. In this paper, we present a methodology to determine a set of 'core classes' by using a range of techniques to reach consensus across several different clustering algorithms, and to ascertain the key characteristics of these classes. We apply the methodology to immunohistochemical data from breast cancer patients. In doing so, we identify six core classes, of which several may be novel sub-groups not previously emphasised in literature.

1. Introduction

Breast cancer, the most common cancer in women [1,2], is a complex disease characterised by multiple molecular alterations. Current routine clinical management relies on availability of robust clinical and pathologic prognostic and predictive factors to support decision making. Recent advances in high-throughput molecular technologies supported the evidence of a biologic heterogeneity of breast cancer.

Following the seminal paper of Eisen and colleagues [3], in which hierarchical clustering and visual inspection of the dendrogram were performed to discover unknown pattern of gene associations, the use of clustering has become more and more popular, especially for discovering profiles in cancer with respect to high-throughput genomic data. Perou et al. [4] identified four molecular distinct breast cancer groups based on gene expression profiles using a hierarchical clustering algorithm: luminal epithelial/estrogen (ER) positive, HER2 positive, basal-like and normal breast-like. A subsequent study extended this by dividing the luminal/ER positive group into three subtypes: luminal-A, B, and C [5], but the luminal-C group was later eliminated [6]. Sotiriou et al. [7] showed six similar groups, with two basal-like subgroups and no normal breast-like group. Whilst numerous studies have reported these and other novel molecular subtypes, and assigned a prognostic significance to the proposed classes [8–10], they remain varied in their detailed classification [11]. An alternative approach to gene expression profiling is to use established robust laboratory technology, such as immunocytochemistry on formalin fixed paraffin embedded patient tumour samples. We and others have applied protein biomarker panels with known relevance to breast cancer, to large numbers of cases using tissue microarrays, exploring the existence and clinical significance of distinct breast cancer classes [12–19]. In particular, in [12] five breast cancer classes were identified and characterised. Note that a sixth group of only four cases was also identified but considered too small for further detailed assessment. However, these studies have not addressed the stability of the proposed classifications across different case sets, assay methods and data analysis procedures. Such an issue appears of critical relevance considering the increase in the number of features involved in bionformatics analyses.

In order to deal with the stability of classifications and in particular of clustering techniques, several studies have focused on the comparison and concordance among different clustering methods defining what is now known as the 'consensus clustering'. Monti and colleagues presented a new methodology of class

* Corresponding author. Tel.: +44 115 95 14229.
E-mail address: dqs@cs.nott.ac.uk (D. Soria).

0010-4825/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2010.01.003

discovery and clustering validation tailored to the task of analysing gene expression data [20]. The new 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. The basic assumption of this method was the following: if the data represent a sample of items drawn from distinct sub-populations, and if a different sample drawn from the same sub-populations were to be observed, the induced cluster composition and number should not be radically different. Therefore, the more the attained clusters are robust to sampling variability, the more one can be confident that these clusters represent real structure.

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 [21]. To assess gene-expression cluster consistency, the use of the weighted-kappa metric was analysed. This metric rates the agreement between the classification decisions made by two or more observers. In this case the two observers are the clustering methods.

Filkov and Skiena proposed a methodology for consensus clustering as an approach to integrating diverse sources of similarity clustered microarray data [22]. They proposed 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. Under reasonable conditions, the consensus cluster should provide additional information to that of the union of individual data analyses. The goals of consensus clustering are to integrate multiple data sets for ease of inspection, and to eliminate the likely noise and incongruencies from the original classifications. In terms of similarity the consensus partition should be close to all given ones, or in terms of distance, it must not be too far from any of them. One way to do this is to find a partition that minimises the distance to all the other partitions. So, given k different partitions, the target one was identified as the consensus partition.

In another approach [23], robust clusters were identified by the implementation of a new algorithm termed 'Clusterfusion'. 'Clusterfusion' 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. Firstly, an agreement matrix was generated with each cell containing the number of agreements amongst methods for clustering together the two variables represented by the indexing row and column indices. This matrix was then used to cluster variables based upon their cluster agreement. In essence, a clustering technique was applied to the clustering results.

The idea of combining and comparing the results of different clustering algorithms is particularly important in order to evaluate the stability of a proposed classification. In this paper, a methodology is presented to evaluate the stability of six breast cancer classes by comparing the clustering solutions provided by different algorithms. In order to address the standard problem of consensus clustering in which the label of classes is arbitrary, a label was assigned using the six clusters characterised in the work of Abd El-Rehim [12], as a reference for the description of our resulting groups.

2. Material and methods

The four-step methodology for elucidating core, stable classes (groups) of data from a complex, multi-dimensional dataset was as follows:

1. A variety of clustering algorithms were run on the data set (see Section 2.1).
2. Where appropriate, the most appropriate number of clusters was investigated by means of cluster validity indices (see Section 2.2).
3. Concordance between clusters, assessed both visually and statistically, was used to guide the formation of stable 'core' classes of data.
4. A variety of methods were utilised to characterise the elucidated core classes.

The methodology was applied to a well-known set of data concerning breast cancer patients [12] (see Section 2.5) in order to obtain core classes. Once these core classes were obtained, the clinical relevance of the corresponding patient groups were investigated by means of associations with related patient data. All statistical analysis was done using *R*, a free software environment for statistical computing and graphics [24].

2.1. Clustering algorithms

Five different algorithms were used for cluster analysis:

(i) Hierarchical (as per our previous study [12]).
(ii) K-means (KM).
(iii) Partitioning around medoids (PAM).
(iv) Adaptive resonance theory (ART).
(v) Fuzzy c-means (FCM).

2.1.1. Hierarchical clustering

The hierarchical clustering algorithm (HCA) begins with all data considered to be in a separate cluster. It then finds the pair of data with the minimum value of some specified distance metric; this pair is then assigned to one cluster. The process continues iteratively until all data are in the same (one) cluster. A conventional hierarchical clustering algorithm (HCA) was utilised, utilising Euclidean distance on the raw (unnormalised) data with all attributes equally weighted.

2.1.2. K-means clustering

The K-means (KM) technique aims to partition the data into K clusters such that the sum of squares from points to the assigned cluster centres is minimised. The algorithm repeatedly moves all cluster centres to the mean of their Voronoi sets (the set of data points which are nearest to the cluster centre). The objective function minimised is

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

where x_i is the i-th datum, v_j is the j-th cluster centre, k is the number of clusters, c_j is the number of data points in the cluster j and $\|x_i - v_j\|$ is the Euclidean distance between x_i and v_j.

The j-th centre v_j can be calculated as

$$v_j = \frac{1}{c_j} \sum_{i=1}^{c_j} x_i, \quad j = 1, \ldots, k$$

K-means clustering is dependent on the initial cluster centres setting (which, in turn, determines the initial cluster assignment). Various techniques have been proposed for the initialisation of clusters [25], but for this study we used a fixed initialisation of the cluster centres obtained with hierarchical clustering. The number of clusters is an explicit input parameter to the K-means algorithm.

2.1.3. Partitioning around medoids

The partitioning around medoids (PAM) algorithm (also known as the k-medoids algorithm) is a technique which attempts to