Information Sciences 262 (2014) 32–45

Contents lists available at ScienceDirect

Information Sciences

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

Analysing microarray expression data through effective clustering

E. Masciari [a,*], G.M. Mazzeo [a], C. Zaniolo [b]

[a] ICAR-CNR, Via P. Bucci 41/C, 87036 Rende, Italy
[b] UCLA, Los Angeles, United States

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form 6 November 2013
Accepted 5 December 2013
Available online 11 December 2013

Keywords:
Bioinformatics
Clustering
Biological data analysis

ABSTRACT

The recent advances in genomic technologies and the availability of large-scale microarray datasets call for the development of advanced data analysis techniques, such as data mining and statistical analysis to cite a few. Among the mining techniques proposed so far, cluster analysis has become a standard method for the analysis of microarray expression data. It can be used both for initial screening of patients and for extraction of disease molecular signatures. Moreover, clustering can be profitably exploited to characterize genes of unknown function and uncover patterns that can be interpreted as indications of the status of cellular processes. Finally, clustering biological data would be useful not only for exploring the data but also for discovering implicit links between the objects. To this end, several clustering approaches have been proposed in order to obtain a good trade-off between accuracy and efficiency of the clustering process. In particular, great attention has been devoted to hierarchical clustering algorithms for their accuracy in unsupervised identification and stratification of groups of similar genes or patients, while, partition based approaches are exploited when fast computations are required. Indeed, it is well known that no existing clustering algorithm completely satisfies both accuracy and efficiency requirements, thus a good clustering algorithm has to be evaluated with respect to some external criteria that are independent from the metric being used to compute clusters. In this paper, we propose a clustering algorithm called M-CLUBS (for Microarray data CLustering Using Binary Splitting) exhibiting higher accuracy than the hierarchical ones proposed so far while allowing a faster computation with respect to partition based approaches. Indeed, M-CLUBS is faster and more accurate than other algorithms, including k-means and its recently proposed refinements, as we will show in the experimental section. The algorithm consists of a divisive phase and an agglomerative phase; during these two phases, the samples are repartitioned using a least quadratic distance criterion possessing unique analytical properties that we exploit to achieve a very fast computation. M-CLUBS derives good clusters without requiring input from users, and it is robust and impervious to noise, while providing better speed and accuracy than methods, such as BIRCH, that are endowed with the same critical properties. Due to the structural feature of microarray data (they are represented as arrays of numeric values), M-CLUBS is suitable for analyzing them since it is designed to perform well for Euclidean distances. In order to stronger the obtained results we interpreted the obtained clusters by a domain expert and the evaluation by quality measures specifically tailored for biological validity assessment.

* Corresponding author. Tel.: +39 0984831735; fax: +39 0984839054.
E-mail addresses: masciari@icar.cnr.it (E. Masciari), gmazzeo@gmail.com (G.M. Mazzeo), zaniolo@cs.ucla.edu (C. Zaniolo).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.12.003

1. Introduction

Nowadays, microarray experiments allow the exploration of huge amounts of gene expressions using a single chip. Moreover, the relatively moderate cost for a chip and the small sample preparation times, enable the analysis of a large number of different experimental conditions, such as points of time-series experiments or disease progression in a cohort of patients [33].

This huge amount of data poses many challenges to the bioinformatics community such as finding the behavior of set of related genes in different conditions. This goal is often achieved by means of cluster analysis, i.e. the identification of similar patterns in different conditions [25]. Indeed, the ability to gather genome-wide expression data has far outstripped the ability of human brains to process the raw data, thus cluster analysis can help scientists to distill the data down to a more comprehensible level by subdividing the genes into a smaller number of categories and then analyzing those [7,9,15].

Further motivation for the exploitation of cluster analysis for biological data lies in the fact that similar patterns found by clustering may correspond to co-regulation of genes [21]. Moreover, cluster analysis represents a fundamental and widely used method of knowledge discovery [26], due to the valuable information it can provide. In particular, the use of cluster analysis has become a standard method in literature for the analysis of microarray expression data used both for initial screening of patients as well as for extraction of molecular signatures of disease [24] or feature selection [5,30]. By cluster analysis, microarray data researcher can focus on finding group of genes that exhibit a similar and coherent evolutionary patterns in a set of patients or time-points. For instance Bayesian approaches have been largely used for data analysis, but their limited scalability and efficiency prevent their use in large scale microarray datasets [27,28,39]. Analogously, a large number of existing algorithms has been applied to microarray data starting from well-known approaches; among those we mention here partition-based clustering (e.g. *k-means* [36]) and its variants (e.g. *fuzzy c-means* [14]), density based clustering (e.g. *DBScan* [17]), hierarchical methods (e.g. *BIRCH* [47], R/BHC [40]), and grid-based methods (e.g. *STING* [44,45]). In particular, agglomerative hierarchical clustering has been used to partition set of patients into smaller groups characterized by exploiting information on set of genes exhibiting similar evolution with respect to a set of similar conditions (e.g. clinical conditions, time evolution or drug responses) [32].

Nevertheless, the logical and algorithmic complexities of this many-facet problem make this research activity quite intriguing. Indeed, in spite of the new progress achieved in recent years (e.g., agglomerative clustering [34], biclustering [1], genetic algorithm based clustering [35], non-metric clustering [19]), significant progress should be expected in the future. In particular, it is well known that no clustering algorithm completely satisfies both accuracy and efficiency requirements, thus a good clustering algorithm has to be evaluated with respect to some external criteria that are independent from the metric being used to compute clusters. As an example, bootstrapping techniques have often been used to calculate the significance of the obtained dendrogram [29].

In this paper, we propose M-CLUBS, a novel algorithm that exhibits quite good performances, in term of *speed, repeatability, accuracy* and *robustness to noise*. M-CLUBS performances have been evaluated using widely accepted clustering validity metric that are method independent thus quite reliable. M-CLUBS excellent performances arise from some key feature of our algorithm, in particular:

- M-CLUBS is not tied to a fixed grid differently from grid-based methods (e.g. *STING* [44]),
- it can backtrack on previously wrong calculation since it performs first a top-down splitting of data and then (eventually) it performs a bottom-up refinement of the obtained results,
- it performs also well on non-globular clusters (i.e. clusters that are not spherical in shape) differently from *k-means* [36] and *BIRCH* [47].

In the following, after a presentation of our method, we show the M-CLUBS properties and finally as proof-of-principle we discuss the performance of our algorithm using some publicly available dataset.

2. Approach

In this paper we propose a new hierarchical algorithm called M-CLUBS (for Microarray data CLustering Using Binary Splitting) whose *speed* performances are better than *k*-means *and* whose *accuracy* overcomes previous hierarchical algorithms while operating in a *completely unsupervised* fashion. The first phase of the algorithm is divisive, as the original data set is split recursively into mini-clusters through successive binary splits: the algorithm's second phase is agglomerative since these mini-clusters are recombined into the final result. Due to its features our algorithm can be used also for refining other approaches performances. As an example it can be used to overcome *k*-means initial assignment problem since its low complexity will not affect the overall complexity while the accuracy of our results will guarantee an excellent initial assignment of cluster centroids. Further, our approach induces during execution a *dynamic hierarchical* grid that will better fit the dataset with respect to classical grid approaches that exploit a fixed grid instead. Finally, the algorithm exploits the analytical properties of the Sum of Squares (SSQ in the following) function to minimize the cost of merge and split operations, and indeed the approach results really fast. One may argue that many different measures could be used for cluster computation but the