Journal of Computational Science xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Computational Science

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

Clustering approaches for dealing with multiple DNA microarray datasets

Veselka Boeva*

Department of Computer Systems and Technology, Technical University of Sofia-Branch Plovdiv, Tsanko Dyustabanov 25, 4000 Plovdiv, Bulgaria

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 27 April 2013
Accepted 7 May 2013
Available online xxx

Keywords:
Consensus clustering
Formal Concept Analysis
Integration analysis
Gene expression data
Particle Swarm Optimization
Partitioning algorithms
Supervised clustering

ABSTRACT

This paper centres on clustering approaches that deal with multiple DNA microarray datasets. Four clustering algorithms for deriving a clustering solution from multiple gene expression matrices studying the same biological phenomenon are considered: two unsupervised cluster techniques based on information integration, a hybrid consensus clustering method combining Particle Swarm Optimization and k-means that can be referred to supervised clustering, and a supervised consensus clustering algorithm enhanced by Formal Concept Analysis (FCA), which initially produces a list of different clustering solutions, one per each experiment and then these solutions are transformed by portioning the cluster centres into a single overlapping partition, which is further analyzed by employing FCA. The four algorithms are evaluated on gene expression time series obtained from a study examining the global cell-cycle control of gene expression in fission yeast *Schizosaccharomyces pombe*.

1. Introduction

Gene clustering is one of the most frequently used analysis methods for DNA microarray data. Clustering algorithms are used to divide genes into groups according to the degree of their expression similarity. Such a grouping may suggest that the respective genes are correlated and/or co-regulated, and moreover that the genes could possibly share a common biological role.

Clustering has traditionally been a tool of unsupervised learning. In unsupervised learning such as clustering, the task is to segment unlabeled training data into clusters that reflect some meaningful structure in the data. Recently, several supervised clustering algorithms have been proposed [1,2,12]. Supervised clustering deviates from traditional clustering in that it is applied on classified examples with the objective of identifying clusters that have high probability density with respect to a single class. Moreover, in supervised clustering, we also like to keep the number of clusters small, and objects are assigned to clusters using a notion of closeness with respect to a given distance function [12].

In this paper, we discuss and compare four methods for clustering of multiple DNA microarray experiments studying the same biological phenomenon. Gene expression microarrays are the most commonly available source of high-throughput biological data. Each microarray experiment is supposed to measure the gene expression levels of a set of genes in a number of different experimental conditions or time points. Microarray experiments are often performed over many months, and samples are often collected and processed at different laboratories. Therefore, it is sensible to think of integrating related results from several microarray studies addressing a similar biological question in order to draw more reliable and robust conclusion. Initially, we consider two unsupervised cluster techniques that are based on information integration [24]. One approach combines the information containing in multiple microarray experiments at the level of expression or distance matrices and then applies a partitioning algorithm on the combined matrix. The second technique aggregates partitioning results derived from multiple microarray data sets. Then we present a consensus clustering method that combines Particle Swarm Optimization and k-means for deriving a global clustering solution for multiple gene expression matrices [7]. It can be referred to supervised clustering since initially the involved datasets are clustered by applying k-means algorithm. Then the final clustering solution is found by updating the cluster centres using the information on the best clustering solution generated by each dataset and the entire set of datasets. We conclude the discussion by revealing the characteristics of a consensus clustering algorithm that is enhanced by Formal Concept Analysis (FCA) [8]. It can also be considered as a supervised clustering technique. However, in contrast to the above interpretation of supervised clustering this approach initially produces a list of different clustering solutions, one per each experiment. These solutions are further transformed by portioning the cluster centres into a single overlapping partition, which is further analyzed by employing FCA [14]. FCA produces a concept lattice where each concept represents a subset of genes that belong to a number of clusters. The concepts compose the final disjoint clustering partition.

* Tel.: +359 895587484.
E-mail address: vboeva@tu-plovdiv.bg

1877-7503/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jocs.2013.05.003

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the clustering methods that deal with multiple microarray datasets. Section 4 introduces the experimental setup. Section 5 shows the validation results and present discussions. Section 6 concludes the paper.

2. Related work

Our goal of this work is to study and compare the performance of four clustering algorithms that derive a clustering solution from multiple gene expression matrices. The combination of data from multiple microarray studies addressing a similar biological question is gaining high importance in the recent years due to the ever increasing number and complexity of the available gene expression datasets [9,15,16,40]. There are several approaches to combine within a clustering process the information contained in different gene representations [19,37]. One representation considered in this case are gene expression data received from a single microarray experiment. The other representation is Gene Ontology containing knowledge about e.g., gene functions gained throughout the years. Studies concerning different ways of combining gene information representations at the level of similarity matrices have been proposed in [19,25]. An approach to combining data from multiple microarray experiments is the aggregation of their clustering into a consensus or representative clustering which increases the confidence in the common features in all the datasets and reveals the important differences among them. Methods for the combination of clustering results derived for each dataset separately have been considered in [13,20,34,36]. For instance, multiple heterogeneous data sets are integrated in [13] by constructing a consensus partition that minimizes the distance to all the other partitions. The algorithm proposed in [20] first generates local cluster models and then combines them into a global cluster model of the data. The study in [36] focuses on clustering ensembles, i.e. seeking a combination of multiple partitions that provides improved overall clustering of the given data. The combined partition is found as a solution to the corresponding maximum likelihood problem using the Expectation-Maximization (EM) algorithm [11]. Strehl and Ghosh consider the problem of combining multiple partitions of a set of objects into a single consolidated clustering without accessing the features or algorithms that determined these partitions [34]. The cluster ensemble problem is formalized as a combinatorial optimization problem in terms of shared mutual information.

The FCA [14] or *concept lattice approach* has been applied for extracting local patterns from microarray data in [3,4] or for performing microarray data comparison in [10,30]. For example, the FCA method proposed in [10] builds a concept lattice from the experimental data together with additional biological information. Each vertex of the lattice corresponds to a subset of genes that are grouped together according to their expression values and some biological information related to the gene function. It is assumed that the lattice structure of the gene sets might reflect biological relationships in the dataset. In [22], a FCA-based method is proposed for extracting groups or classes of co-expressed genes. A concept lattice is constructed where each concept represents a set of co-expressed genes in a number of situations. A serious drawback of the method is the fact that the expression matrix is transformed into a binary table (the input for the FCA step) which leads to possible introduction of biases or information loss.

3. Techniques for clustering of multiple microarray data sets

We will now review four algorithms for clustering of multiple microarray experiments studying the same biological phenomenon by first discussing two unsupervised clustering techniques based on information integration followed by a consideration of a hybrid clustering approach combining Particle Swarm Optimization and k-means clustering and then by revealing the advantages of a consensus clustering algorithm that is enhanced by Formal Concept Analysis. Notice that the first algorithm combines the information contained in multiple related microarray experiments at the level of expression or distance matrices and then applies a partitioning algorithm on the combined matrix while the other three are consensus clustering techniques that integrate partitioning results derived separately from the microarray experiments. In addition, the last two algorithms can be referred to supervised clustering.

3.1. Clustering of multiple microarray experiments using information integration

Kostadinova et al. have shown in [24] how two microarray data integration techniques [5,39] can be applied to both definitions of the problem of deriving clustering results from a set of gene expression matrices: (1) information contained in different data sets may be combined at the level of expression (or similarity) matrices and then cluster; (2) given multiple clusterings, one per each data set, find a combined (consensus) clustering solution.

First, let us consider a cluster integration approach, proposed in [24], which combines the information contained in multiple microarray experiments at the level of expression or distance matrices and then applies a partitioning algorithm on the combined matrix. Assume that a particular biological phenomenon is monitored in a few high-throughput experiments under n different conditions. Each experiment i ($i = 1, 2, \ldots, n$) is supposed to measure the gene expression levels of m_i genes in n_i different experimental conditions or time points. Thus, a set of n different data matrices M_1, M_2, …, M_n will be produced, one per experiment. Initially, the set of studied genes is restricted to those contained into all datasets, i.e. a set of m overlapping genes is found across all datasets. Initially, some integration procedure (hybrid integration or hierarchical merge) is applied to transform the set of input matrices M_1, M_2, …, M_n into a single matrix, which values can be interpreted as consensus values supported by all the experiments. Then, the overall matrix is passed to the corresponding clustering algorithm for subsequent analysis. In [24], this idea is demonstrated by implementing two partitioning algorithms (see Appendix A): k-medoids and k-means. Since the k-medoids clustering algorithm is suitable for cases in which the distance matrix is known but the original data matrix is not available, the hybrid integration procedure [5] is used to combine the quadratic distance (similarity) matrices, generated per each considered data set. On the other hand, k-means clustering method requires an original expression data matrix as input data set and thus, the hierarchical merge algorithm [39] is used to merge the expression profiles from the original input matrices. Finally, the obtained integrated similarity (or fused expression) matrix is passed to k-medoids (or k-means) clustering algorithm for subsequent analysis. Notice that, by applying information about the quality of the microarrays, weights may be assigned to the experiments and can be further used in the integration process in order to obtain more realistic overall values.

In [24], Kostadinova et al. have also introduced a consensus clustering algorithm, referred to *Integrative* clustering, that integrates partitioning results derived from multiple microarray data sets. Initially, k cluster centres are initialized by using the information contained in the studied datasets in an integrated manner. The selected partitioning algorithm can then be applied to each expression matrix, which will generate a set of partition matrices: $P_1, \ldots, P_n$. Each partition matrix may be represented as $P_r = \{p^r_{ij}\}$, where p^r_{ij} is the membership of gene j ($j = 1, \ldots, m$) to the ith ($i = 1, \ldots, k$)

Please cite this article in press as: V. Boeva, Clustering approaches for dealing with multiple DNA microarray datasets, J. Comput. Sci. (2013), http://dx.doi.org/10.1016/j.jocs.2013.05.003