A Latent Feature Model Approach to Biclustering

José Caldas, Aalto University, Espoo, Finland
Samuel Kaski, Aalto University, Espoo, Finland

ABSTRACT

Biclustering is the unsupervised learning task of mining a data matrix for useful submatrices, for instance groups of genes that are co-expressed under particular biological conditions. As these submatrices are expected to partly overlap, a significant challenge in biclustering is to develop methods that are able to detect overlapping biclusters. The authors propose a probabilistic mixture modelling framework for biclustering biological data that lends itself to various data types and allows biclusters to overlap. Their framework is akin to the latent feature and mixture-of-experts model families, with inference and parameter estimation being performed via a variational expectation-maximization algorithm. The model compares favorably with competing approaches, both in a binary DNA copy number variation data set and in a miRNA expression data set, indicating that it may potentially be used as a general-problem solving tool in biclustering.

KEYWORDS

Biclustering, Bioinformatics, Gene Expression, Genomics, Machine Learning

INTRODUCTION

Clustering methods have been useful at understanding global trends in high-throughput molecular biology data (D’Haeseleer, 2005). Due to the increasing number of phenotypes that are probed for in individual studies, as well as the intrinsic complexity and high dimensionality of genome-wide high-throughput data, it has become increasingly meaningful to detect local, rather than global data trends. In practice, this amounts to grouping subsets of biological conditions and associate with each group the measurements that make those biological conditions similar. For instance, in a gene expression study where the messenger RNA (mRNA) level of each gene is measured in a number of samples corresponding to certain biological conditions (e.g. tumor samples from multiple disease stages), the analyst may be interested in detecting groups of samples that are similar and associate with each group of samples the genes whose measurements are similar across those samples. This unsupervised learning task is known as biclustering (Cheng & Church, 2000; Madeira & Oliveira, 2004).

Formally, biclustering is an unsupervised learning task that takes as input a data matrix $D$ and learns a set of submatrices of $D$, which are designated as biclusters. Each submatrix/bicluster should contain certain desirable properties; the specific desiderata depend on the specific problem formulation and the data type of the input matrix $D$. For instance, in a sparse binary matrix, the analyst’s intention may be to detect biclusters that correspond to dense submatrices; alternatively, in a continuous data set with values spanning a given range, the modeller’s intention may be to detect biclusters that correspond to submatrices in which the values are typically close to each other according to a given metric (e.g. Euclidean). Throughout the present paper, we use the terms object and condition to
designate the rows and columns of a data matrix. For instance, in a gene expression matrix such as the one described above, the objects are genes and the conditions are the biological samples.

Biclustering methods may be classified according to the type of bicluster structures they can detect (Madeira & Oliveira, 2004). Crucially, some but not all methods allow biclusters to overlap, that is, they allow a pair \( (i, j) \) of object \( i \) and condition \( j \) to belong to more than one bicluster. Allowing membership to multiple biclusters is justified, accounting for instance to the multiple functional roles that a gene may undertake or the various biological processes that are simultaneously active in a biological condition. However, it brings in additional technical challenges regarding how to properly specify a model that handles bicluster overlap. A typical approach in the context of expression data is to specify a linear model wherein each bicluster corresponds to a given set of parameters; in this model family, bicluster parameters combine additively, \( i.e., \) the parameters used for modelling a given data point \( y_{ij} \) are obtained by summing the parameters across all biclusters that include the object-condition pair \( (i, j) \). A well-known member of this model family is the plaid model (Lazzeroni & Owen, 2002). More general frameworks combine parameter additivity with link functions \( i.e., \) the sigmoid function \( \) in order to model discrete data types (Meeds et al., 2007). The main drawbacks of such parameter interaction paradigms are that the specific interaction assumptions are restrictive and often may not hold; the necessity to introduce parameter interaction assumptions, \( e.g., \) additive combination of bicluster parameters, in addition to specifying how each bicluster models the data assigned to it, may lead the practitioner to postulate artificial assumptions solely for the purpose of maintaining model soundness. We propose an alternative mixture-modelling approach, leading to more straightforward solutions, in which each object-condition pair \( (i, j) \) may belong to several biclusters, as long as those biclusters provide equally good models for the corresponding data points \( y_{ij} \).

In this paper we introduce a biclustering framework which implements the notion that membership of an object-condition pair \( (i, j) \) in multiple biclusters depends on how well the biclusters independently model the corresponding data point \( y_{ij} \). This approach does not require making an analytical specification of how the parameters corresponding to each bicluster combine at the overlapping regions. Our approach can be conceptually regarded as belonging to the mixture-of-experts model family (Bishop & Svensén, 2003; Jordan & Jacobs, 1994). Mixture-of-experts models assume the existence of a certain number of data generators, or experts, with each expert consisting of a relatively simple data model that is typically adequate only for a part of the data set. In a data set consisting of several input-output pairs \( (x, y_i) \), \( e.g., \) in classification or regression data sets, the input \( x_i \) probabilistically determines which expert is used to model the output \( y_i \). In the current context, data points are rather of the form \( ((i, j), y_{ij}) \), \( i.e., \) the input consists of indices describing which object \( i \) and condition \( j \) the output \( y_{ij} \) corresponds to. The set of triplets \( ((i, j), y_{ij}) \) across all rows \( i \) and columns \( j \) is simply another way of encoding the input data matrix. Our proposed framework allows the contextual information \( (i, j) \) to probabilistically determine which expert is used to model \( y_{ij} \). In detail, we let each bicluster specify a so-called region of expertise, consisting of a group of objects and conditions whose corresponding data points are adequately modeled by the bicluster. By making use of a flexible latent feature model prior distribution, these regions of expertise may freely overlap; the membership of an object-condition pair \( (i, j) \) to each of these regions of expertise determines the probability of choosing a given bicluster to model the corresponding data point \( y_{ij} \). Strictly speaking, data points do not belong to multiple biclusters; there is instead uncertainty over which bicluster each data point belongs to. However, the regions of expertise inferred by the model that connect the contextual information \( (i, j) \) to the probability of choosing each bicluster effectively yield overlapping bicluster structures. We apply the framework to binary DNA copy number variation (CNV) data and continuous miRNA expression data, showing that the framework performs well and lends itself to multiple data types and model assumptions.
Clustering Genes Using Heterogeneous Data Sources
www.igi-global.com/article/clustering-genes-using-heterogeneous-data/45163?camid=4v1a