2010 International Conference on Pattern Recognition

Biclustering of expression microarray data with topic models

Manuele Bicego*†, Pietro Lovato*, Alberto Ferrarini* Massimo Delledonne*
*University of Verona, Verona, Italy 37134
Contact email: manuele.bicego@univr.it
†Istituto Italiano di Tecnologia (IIT), Genova, Italy

***Abstract*—This paper presents an approach to extract biclusters from expression microarray data using topic models – a class of probabilistic models which allow to detect interpretable groups of highly correlated genes and samples. Starting from a topic model learned from the expression matrix, some automatic rules to extract biclusters are presented, which overcome the drawbacks of previous approaches. The methodology has been positively tested with synthetic benchmarks, as well as with a real experiment involving two different species of grape plants (*Vitis vinifera* and *Vitis riparia*).**

***Keywords*-Expression Microarray, graphical models, topic model, biclustering**

I. Introduction

Pattern Recognition techniques have shown to be very useful in the analysis of expression microarray data [1], for example for classification and clustering. In the clustering context, a recent trend is represented by the application of biclustering methodologies, namely clustering techniques able to simultaneously group genes and samples [2], [3]. Different approaches have been presented in the literature in the past, each one characterized by different features, like computational complexity, effectiveness, interpretability, optimization criterion and others – for a review see [2], [3].

This paper approaches the expression microarray biclustering problem with a particular class of statistical models, typically known as *topic* or *latent models* – e.g. the Probabilistic Latent Semantic Analysis (PLSA – [4]). These powerful approaches have been introduced in the text understanding community for unsupervised topic discovery in a corpus of documents, and subsequently largely applied in the computer vision community [5]. One of the main characteristics of this class of approaches is represented by their interpretability. Actually they can model a dataset in terms of hidden topics (or processes), which can reflect underlying and meaningful structures in the problem. In the expression microarray context, such models have been used for clustering [6] and for classification [7]. In the context of the biclustering, only one quite recent paper [8] applied topic models to microarray data. In particular, in that paper the goal was to elucidate the cell type-specific transcription factors, using genomic sequences as well as expression profiles. Nevertheless, the method used there to characterize biclusters, given the PLSA model, is rather simplistic, having different drawbacks, like assigning the same number of genes to every bicluster and not permitting overlap towards samples (namely each sample may belong to only one bicluster). In this paper we go one step forward in this direction, proposing some simple but powerful rules which permit to extract overlapped clusters with a reasonable number of genes. When possible, automatic methods for choosing the parameters have been proposed. The proposed rules have been favorably compared with [8] using the synthetic data and protocols described in [3]. Moreover, the proposed approach has been tested on a real experiment, involving two different species of grape plants (*Vitis vinifera* and *Vitis riparia*). A qualitative validation, involving Gene Ontology and a priori biological knowledge, confirms the appropriateness of the proposed approach.

II. Topic Models

Topic models were introduced in the linguistic scenario, in order to describe and model documents. The basic idea underlying these methods is that each document is characterized by the presence of one or more topics (e.g. sport, finance, politics), which may induce the presence of some particular words. From a probabilistic point of view, the document may be seen as a mixture of topics, each one providing a probability distribution over words. It is possible to infer the set of topics that were responsible for generating a collection of documents.

In the following we will briefly review the mathematics of the topic model employed in this paper, namely PLSA.

In the PLSA [4] the input is a dataset of N documents $\{d_i\}, i = 1,...,N$, each one containing a set of words. Before applying PLSA, the dataset is summarized by a co-occurrence matrix of size $M \times N$, where the entry $<w_j, d_i>$ indicates the number of occurrences of the word w_j in the document d_i, also called $n(w_j, d_i)$. Each document d_i has n_i words. The presence of a word w_j in the document d_i is mediated by a latent *topic* variable, $z \in Z = \{z_1,..., z_Z\}$, also called *aspect* class, *i.e.*,

$$P(w_j, d_i) = \sum_{k=1}^{Z} P(w_j|z_k)P(z_k|d_i)P(d_i). \qquad (1)$$

The hidden distributions of the model, $P(w|z)$ and $P(z|d)$, are learnt using Expectation-Maximization (EM) – for a deeper review of PLSA, see [4].

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.668

III. The proposed approach

An expression microarray experiment measures the expression level of a set of genes in a pool of biological samples. As a result, an expression matrix is derived, where each row represents the expression level of a gene among all samples, and each column represents the expressions of all genes for a particular sample.

Topic models may be very useful in the expression microarray context, since they may provide powerful and interpretable descriptions of experiments. In particular there is an analogy between the pairs *word-document* and *gene-sample*: actually it seems reasonable to intend the samples as documents and the genes as words. In fact the expression level of a gene in a sample may be easily interpreted as the count of words in a document (the higher the number the more present/expressed the word/gene is). In our case, therefore, we can consider the expression matrix as the count matrix $< w_j, d_i >$ of topic models, after a proper normalization in order to have positive and integer values.

A. The relation topic/bicluster.

In our context, a topic may characterize a subset of samples where the gene expressions are highly correlated. This is exactly the concept of a bicluster: therefore it is reasonable to assume that each topic characterizes a particular bicluster, which may correspond to a particular biological process. This representation is highly informative: the probability $P(w|z)$ may be interpreted as the impact of the different genes in a particular biological process. On the other side, $P(z|d)$ may be used to infer the different biological processes which are active over the different samples. Moreover, the probabilistic nature of these models permit to encode also the *level* of the impact: in biology, it is known that not all genes involved in a biological process have the same importance or the same impact on it; on the other side, not all biological processes are involved in every sample.

Even if this probabilistic membership of each gene/sample to a particular bicluster may be really advantageous (leading to the concept of *soft biclusters*), an automatic mechanism able to explicitly list the components of a bicluster may be useful as well, especially for validation purposes (to the best of our knowledge, no biological validation tools deal with probabilistic memberships).

B. From topics to hard biclusters.

In order to crisply define the content of a bicluster, in [8] the authors assign each sample d to the topic z for which the probability $P(z|d)$ is maximum; then, for each bicluster, genes are selected by retaining only a percentage of genes sorted by descending probability $P(w|z)$ (they assumed that 7% was a good percentage). Clearly this methodology is rather simplistic, since it assumes that each sample can belong to only one bicluster and that each bicluster is composed by the same number of genes (independently from the level of relevance of such genes). In this paper we investigate some more sophisticated rules, which permit to have more significant biclusters. These rules can be applied either to extract documents or to extract genes, simultaneously or in combination, starting from the PLSA distributions $P(w_j|z)$ and $P(z|d_i)$.

1) Max Rule.: In this case a gene w_j (or the sample d_i) is assigned to the bicluster showing the highest $P(w_j|z)$ (or $P(z|d_i)$). This is the rule employed in [8] to assign samples to the biclusters. Clearly, in this way, we have an exhaustive assignment of genes and samples to the biclusters: this may imply that also genes (or samples) which are irrelevant – namely with a low $P(w|z)$ (or $P(z|w)$) – may be assigned to a bicluster. Secondly, this rule does not manage overlapping biclusters: in other words genes and samples can not belong to different biclusters, this being a severe limitation in this biological context: assuming that a bicluster (a topic) describes a particular biological process, it is highly possible that the same gene participates to different biological processes, with different roles; on the contrary, the same biological process may be involved in different samples.

2) Percentage Rule.: In this case, for each bicluster, only the $X\%$ most probable elements are retained. This rule permits overlapping biclusters, but assigns to each bicluster the same number of genes (samples), possibly including also genes or samples which are not relevant – namely with a low $P(w|z)$ (or $P(z|w)$). Another drawback of this method is that a proper value X should be chosen. This is the rule employed in [8] to assign genes to the biclusters – in that paper X was set to 7.

3) Threshold Rule.: With this rule, only significant genes (samples) are assigned to each bicluster, namely only those genes (samples) whose probability $P(w|z)$ (or $P(z|d)$) is above a given threshold θ. This permits to obtain biclusters which are overlapped, containing only significant genes and samples. Of course the choice of θ is crucial, and depends on the particular dataset under investigation. It can be set on the basis of the biological a priori knowledge, or it may be estimated from the data. Here we tested a series of reasonable choices, based on the mean (or the median) of the probabilities $P(w|z)$ (or $P(z|d)$). In particular, we investigated the effectiveness of θ equal to the mean, the mean plus the standard deviation, mean plus twice the standard deviation, together with their robust estimates (namely median, median plus $5/3$ of median absolute deviation (mad), median plus $10/3$ of mad) [9].

4) GMM rule.: This rule derives from the observation that in an ideal case the pdf of the $P(w|z)$ (or $P(z|d)$) values should assume a bimodal behavior, with one mode for not important elements and one for the active ones. Following this intuition, in this rule we used Gaussian Mixtures (with

2 Gaussians) to automatically model the probability $P(w|z)$ (or $P(z|d)$), using the posterior probability to classify each gene (sample), retaining as relevant only the elements assigned to the "relevant" mode.

IV. Experiments

The methodology proposed in this paper has been tested on synthetic and real datasets, as detailed in the following.

A. Synthetic evaluation.

We tested all different rules in a synthetic benchmark ([3]), which includes synthetic expression matrices, perturbed with different levels of noise and different levels of overlap[1]. In the former case, the expression matrix has been perturbed with gaussian noise, for an increasing variance. In the latter case, the biclusters are overlapped, with different degrees of overlap. The accuracy of the biclustering has been assessed with the so-called *Gene Match Score* [3], which reflects the similarity of the biclusters obtained by an algorithm and the original known biclustering (it varies between 0 and 1 – the higher the better the accuracy) – for all details on the datasets and the evaluation protocol please refer to [3].

For each expression matrix, PLSA training was repeated 30 times (and results averaged), the number of topics was set to the number of biclusters (which is known). Since the *Gene Match Score* considers the accuracy only in one direction, here we performed the evaluation varying the rules used to extract the genes. Obtained results are shown in Fig. 1. In particular, all the four rules have been applied. For the threshold rule, the best automatic way to choose the threshold was the mean + standard deviation approach, which is reported in the plot. It is evident that the proposed rules are robust with respect to noise and significantly outperform the rule proposed in [8] (which is the Percentage Rule). Concerning the overlap, Max and Percentage rules are not robust (by definition) to the overlap, thus resulting in poor performances. Obtained results are also competitive to the results presented in [3] (not reported here due to lack of space).

B. Real evaluation.

The proposed approach has been tested with a real dataset, including 52 samples (and 24676 genes) of microarray expressions of leaves and berries of two grape plants (*Vitis vinifera* and *Vitis riparia*), subjected to pathogen infiltration at different time steps[2]. PLSA was trained starting from the expression matrix (properly scaled and normalized); the number of topics, given the known a priori knowledge, was set to 10. We used the Threshold Rule (θ = mean + std), for both the genes and the samples.

[1]All datasets may be downloaded from: www.tik.ee.ethz.ch/sop/bimax.

[2]The full description of the dataset is in a publication currently under review. In case, more details will be given in the camera ready version.

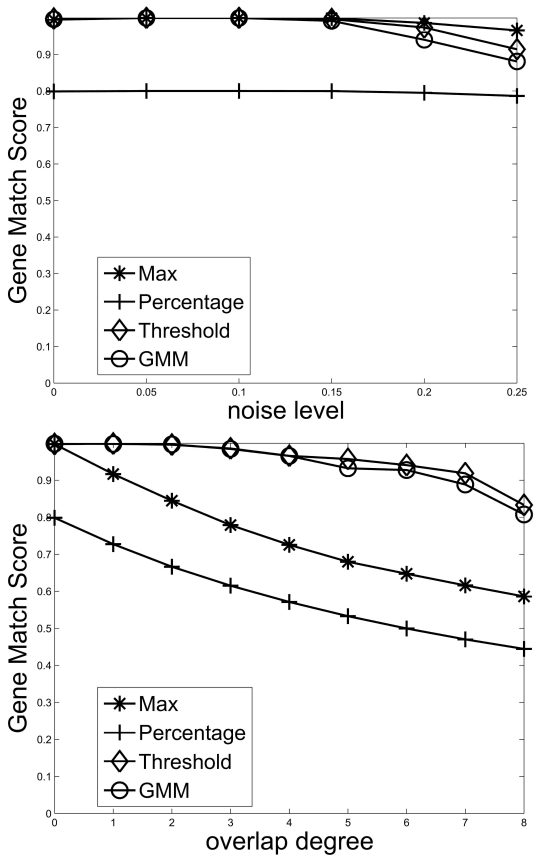

Figure 1. Results on the synthetic dataset: (left) varying noise level; (right) varying overlapping ratio.

In a real case the validation may not be absolute, and should be carried out by employing a priori information on the dataset and on the application scenario. Concerning the samples, instead of listing the obtained biclusters, we report in figure 2 the more intuitive and interpretable $P(z|d)$, from which the samples may be assigned to the biclusters. Studying the composition of the dataset, we observed that

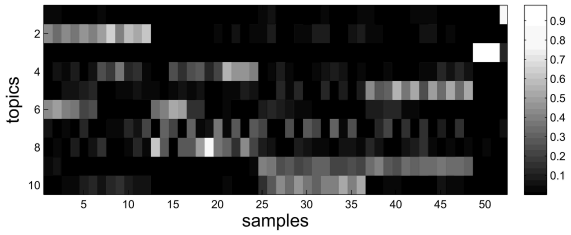

Figure 2. $P(z|d)$ for the real experiment.

it is rather accurately reflected in the probability matrix. In particular, the 52th sample derives from a RNA pool of berries at different growth stages, without any pathogen infiltration, thus representing a single class. This is evidently

captured by the first bicluster (topic), in which the 52th sample is the only one very active – i.e. whose $P(z|d)$ is very high. The samples 48-51 represent another class, containing three technical repetitions derived from a berry of *Vitis vinifera*: this class may be clearly found in the third bicluster. The samples 1-48 represent a more complex class, related to leaves of two particular grape's plants: *Vitis vinifera* and *Vitis riparia*, subjected to pathogen infiltration at different time steps. This part has been chosen because it contains different overlapped sub classes: actually the samples can be divided on the basis of the type of grape plant, of the time of development, and of the type of infiltration. In particular, using the first two criteria, the classes are 1-12, 13-24, 25-36 and 26-48. This is clearly reflected in the biclusters (topics) 2, 4, 10 and 5. If we choose another criterion – like the first and the third – we may have other groups, visible for example in the bicluster 6. From these observations it seems that the PLSA is able to characterize the different groups in the datasets; moreover it is evident the need of rules able to manage also overlaps (contrarily to the one proposed in [8]).

An evaluation of the genes extracted from our approach has been carried out. In particular, we determined the over-represented Gene Ontology [10] terms in the biclusters using the GOstat tool[3], looking for correlation of terms to processes which are known in the specific context. For example, in the 6th bicluster, we have found terms which are related to primary metabolism, response to stimulus and response to stress. Induction of genes belonging to these classes as an effect of pathogen infiltration is in line with the biological knowledge of the mechanisms which regulate the interaction between plants and pathogens, which may lead to a re-programming of the primary metabolism needed to sustain, from an energetic point of view, the organism's response to the pathogen. Even if a further quantitative validation is needed (and currently under investigation), the obtained preliminary findings suggest the suitability of the proposed approach to discover biological information in microarray data.

V. Conclusions

In this paper we presented an approach to extract biclusters from expression microarray data using PLSA. Some automatic rules to extract biclusters from PLSA are presented, which overcome the drawbacks of previous approaches. The methodology has been positively tested with synthetic benchmarks, as well as with a real experiment.

Acknowledgment

M. Bicego acknowledges financial support from the FET programme within the EU FP7, under the SIMBAD project (Contract 213250).

[3]Available at gostat.wehi.edu.au.

References

[1] F. Valafar, "Pattern recognition techniques in microarray data analysis: A survey," *Annals of the New York Academy of Sciences*, vol. 980, pp. 41–64, 2002.

[2] S. Madeira and A. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE Trans. on Computational Biology and Bioinformatics*, vol. 1, pp. 24–44, 2004.

[3] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[4] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1-2, pp. 177–196, 2001.

[5] M. Cristani, A. Perina, U. Castellani, and V. Murino, "Geo-located image analysis using latent representations," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008*, 2008, pp. 1–8.

[6] S. Rogers, M. Girolami, C. Campbell, and R. Breitling, "The latent process decomposition of cdna microarray data sets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 143–156, 2005.

[7] M. Bicego, P. Lovato, B. Oliboni, and A. Perina, "Expression microarray classification using topic models," in *ACM SAC - Bioinformatics and Computational Biology track*, 2010.

[8] J.-G. Joung, D.-H. Shin, R. Seong, and B.-T. Zhang, "Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation," *Bioinformatics*, vol. 22, no. 16, pp. 2005–2011, 2006.

[9] F. Hampel, P. Rousseeuw, E. Ronchetti, and W. Stahel, *Robust Statistics: the Approach Based on Influence Functions*. John Wiley & Sons, 1986.

[10] The Gene Ontology Consortium, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.