Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Traveling on discrete embeddings of gene expression

Pietro Lovato[a,*], Manuele Bicego[a], Maria Kesa[b], Nebojsa Jojic[c], Vittorio Murino[d], Alessandro Perina[c]

[a] Department of Computer Science, University of Verona, Strada le Grazie 15, 37134 Verona, Italy
[b] Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
[c] Microsoft Research, One Microsoft Way, 98052 Redmond, WA, USA
[d] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163 Genova, Italy

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form 20 May 2016
Accepted 21 May 2016

Keywords:
Data visualization
Bag of words clustering
Bag of words classification
Gene selection
Gene expression matrix

ABSTRACT

Objective: High-throughput technologies have generated an unprecedented amount of high-dimensional gene expression data. Algorithmic approaches could be extremely useful to distill information and derive compact interpretable representations of the statistical patterns present in the data. This paper proposes a mining approach to extract an informative representation of gene expression profiles based on a generative model called the Counting Grid (CG).

Method: Using the CG model, gene expression values are arranged on a discrete grid, learned in a way that "similar" co-expression patterns are arranged in close proximity, thus resulting in an intuitive visualization of the dataset. More than this, the model permits to identify the genes that distinguish between classes (e.g. different types of cancer). Finally, each sample can be characterized with a discriminative signature – extracted from the model – that can be effectively employed for classification.

Results: A thorough evaluation on several gene expression datasets demonstrate the suitability of the proposed approach from a twofold perspective: numerically, we reached state-of-the-art classification accuracies on 5 datasets out of 7, and similar results when the approach is tested in a gene selection setting (with a stability always above 0.87); clinically, by confirming that many of the genes highlighted by the model as significant play also a key role for cancer biology.

Conclusion: The proposed framework can be successfully exploited to meaningfully visualize the samples; detect medically relevant genes; properly classify samples.

1. Introduction

Technologies such as gene expression microarrays and RNA-seq provide scientists with a way to measure the expression levels of thousands of genes simultaneously. Computational approaches are increasingly needed to manage this amount of data, and are effectively helping researchers to unravel the complexity of biological systems. Examples of computational problems related to the analysis of a gene expression matrix (a matrix containing the expression level of different genes under different experimental conditions) are classification of samples [1–4], clustering of genes or pathological subtypes [5,6], and selection of differentially expressed or discriminative genes [7].

Other than sophisticated methods of quantitative analysis, high-throughput experiments brought also the need for visualization, thoughtful validation, and, more generally, a deeper understanding of the phenomenon under investigation. For these reasons, interpretable models are required. In this context, generative models (in particular, topic models and latent process models [8]) have been shown to provide highly interpretable solutions, more than achieving high accuracy for classification tasks [9,10]. Within this literature, topic models have been either designed ad hoc for gene expression analysis [11,12], or exported from Natural Language Processing by postulating an analogy between textual documents and microarray samples [10,13]. In the latter case, the starting point is to see a gene expression profile (i.e. a sample) as a "bag of words" vector [14] – a numerical vector in which every entry counts how many times each "word" of a pre-defined dictionary occurs in the considered document. Similarly to text documents, a gene expression profile can be seen as a bag of words vector – genes now represent the words – since each entry measures the

[*] Corresponding author. Tel.: +39 0458027804; fax: +39 0458027068.
E-mail address: pietro.lovato@univr.it (P. Lovato).

http://dx.doi.org/10.1016/j.artmed.2016.05.002
0933-3657/© 2016 Elsevier B.V. All rights reserved.

intensity of expression of each gene (which indirectly reflects the amount of mRNA transcripts). This analogy also permits to exploit topic models in this context [10,13], which, by introducing the concept of "topic", allow to model co-occurrence (or co-expression) patterns within the data. Topics are latent distributions that assign high probability to co-occurring "words", and act as intermediate descriptors of samples (in the gene expression case, they can be associated to biological processes, as shown in [10,13]).

However, a common assumption of most topic models is that the topics act independently of each other. While this assumption is often needed to simplify computations and inference, it may be too simplistic in the gene expression scenario, where it is known that biological processes are tightly co-regulated and interdependent in a complex way. In this paper we make a step forward along this research line – pursuing the topic model philosophy, but coping with the afore-described limitation – presenting a novel strategy to extract an informative representation for a set of experimental samples through a recent generative model called Counting Grid (CG – [15]). The Counting Grid represents a probabilistic model for objects represented as "bag of words", that was recently introduced for text mining [15] and image processing [16]. The idea behind the model is that the topics are arranged on a discrete grid, learned in a way that "similar" topics are closely arranged. Similar biological samples, i.e. sharing topics and active genes, are mapped close on the grid, allowing for an intuitive visualization of the data set. More specifically, the CG seems to be very suitable in the gene expression scenario for the following reasons:

- The CG provides a powerful representation, which permits to capture evolution of patterns in experiments, and can be clearly visualized.
- The CG is well suited for data that exhibit smooth variation between samples. Expression values are biologically constrained to lie within certain bounds by purifying selection [17] and variation in only a few expression values can cause a pathology. This specific property of the data is captured well by the model.
- The CG permits a principled and founded way to extract the most relevant genes that are associated with a disease [18].
- Last, but not least, it is possible to achieve a better classification accuracy with respect to other topic model approaches, as well as to the recent state of the art.

In this paper, we comprehensively evaluate the CG model for mining and modeling gene expression data; we start from the preliminary findings which appeared in the literature [18,19], but we thoroughly evaluate the capabilities of the model with respect to the following novel aspects:

1. By visualizing different data sets, we show that samples belonging to different biological conditions (such as different types of cancer) cluster together on the grid, supporting this claim with a numerical validation (Section 4.1).
2. We systematically tested the accuracy of the CG model both in a gene selection and in a classification setting, experimenting on 7 different benchmark datasets, obtaining results comparable with the recent state-of-the-art.
3. We prove that the model is able to highlight genes that are involved in the pathology or in the phenomenon which motivated the experiment; moreover, the selected genes have a beneficial effect when used for classification, quantitatively comparable with other gene selection techniques.
4. We evaluate the sensitivity of the model to parameters such as grid and window size and the robustness of the model to overfitting.

2. Methods

2.1. The Counting Grid model

In machine learning research, a data point is often represented as a "bag of words": the representation is obtained by counting how many times each "word" (i.e. constituting feature) occurs in the object. This paradigm can represent in a vector space many types of objects, even ones that are non-vectorial in nature. However, one drawback is that in some domains and applications it destroys the possible structure of objects. A clear example can be found in the Natural Language Processing domain (where the bag of words has been originally introduced): by representing a document as a vector of word counts, the ordering of the words in such document is lost.

Recently, an analogy has been established between the Natural Language Processing and the gene expression contexts [10]: the idea is to directly interpret the gene expression matrix as a bag of words, where genes represent words and a sample $\mathbf{s}^t$ represents a document. The expression value can be seen as a count: the higher the expression, the higher the number of transcripts that will be translated into fully functional proteins. In the past, such bag of words representation of gene expressions has been successfully modeled with topic models [10]: these models, introduced in the text mining community, learn a small number of topics which correlate related genes particularly active in a subset of samples. However, there are no strong constraints in how topics are mixed, because they are assumed to be statistically independent. This is a strong drawback, overcame in the Counting Grid model by arranging these distributions representing topics on a discrete grid with topological constraints: intuitively, similar "topics" are located nearby on the grid, and have similar genes' distributions.

Formally, the Counting Grid $\pi_{\mathbf{i},z}$ is a D-dimensional discrete grid, of size $\mathbf{E}=(E_1, \ldots, E_D)$. Each position on the grid is indexed by $\mathbf{i}=(i_1, \ldots, i_D)$, where $i_d \in \{1, \ldots, E_d\}$. Each cell represents a tight distribution over genes (indexed by z), so $\sum_z \pi_{\mathbf{i},z}=1$. A given sample $\mathbf{s}^t$, represented by expression values $\{g_z^t\}$ is assumed to follow a distribution found in a *window* of dimensions $\mathbf{W}=(W_1, \ldots, W_D)$ somewhere in the counting grid. The window is identified by the location $\mathbf{k}$ (upper-left corner of the window) and includes the grid region $W_\mathbf{k}=[\mathbf{k} \ldots \mathbf{k}+\mathbf{W}]$, that is the region starting from the location $\mathbf{k}$ (upper-left corner of the window) and extending in each direction d by W_d grid positions. For example, in Fig. 1 we show a bidimensional CG containing 10 × 10 cells ($\mathbf{E}=(10, 10)$), where the window has size $\mathbf{W}=(3, 3)$. Assuming that the sample $\mathbf{s}^t$ is generated from the window which starts in position $\mathbf{k}=(3, 8)$, the distribution of its genes is defined as the average of all the distributions from $\pi_{(3,8),z}$ to $\pi_{(5,10),z}$ (zoomed in the right part of Fig. 1). Mathematically, this average – given a gene indexed by z in sample $\mathbf{s}^t$ – is computed as:

$$h_{\mathbf{k},z} = \frac{1}{\prod_d W_d} \sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z} \tag{1}$$

A consequence of this can be seen in Fig. 2: if we consider two samples located nearby (s^1 and s^2 in the figure), we note that they share some cells on the grid, and for this reason their genes' distributions will be similar. In other words, spatial proximity implies similarity of expression values.

More formally, the position (upper-left corner) of the window $\mathbf{k}$ in the grid is a latent variable, given which the probability of the bag of words $\{g_z^t\}$ for sample $\mathbf{s}^t$ is

$$p(\{g_z^t\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{g_z^t} = \left(\frac{1}{\prod_d W_d}\right) \prod_z \left(\sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z}\right)^{g_z^t} \tag{2}$$