Biologically-aware Latent Dirichlet Allocation (BaLDA) for the Classification of Expression Microarray

Alessandro Perina[1], Pietro Lovato[1],
Vittorio Murino[1,2], and Manuele Bicego[1,2,*]

[1] University of Verona, Verona, Italy
[2] Italian Institute of Technology (IIT), Genova, Italy
Tel.: +39 045 8027072, Fax: +39 045 8027968
manuele.bicego@univr.it

Abstract. Topic models have recently shown to be really useful tools for the analysis of microarray experiments. In particular they have been successfully applied to gene clustering and, very recently, also to samples classification. In this latter case, nevertheless, the basic assumption of functional independence between genes is limiting, since many other a priori information about genes' interactions may be available (co-regulation, spatial proximity or other a priori knowledge). In this paper a novel topic model is proposed, which enriches and extends the Latent Dirichlet Allocation (LDA) model by integrating such dependencies, encoded in a categorization of genes. The proposed topic model is used to derive a highly informative and discriminant representation for microarray experiments. Its usefulness, in comparison with standard topic models, has been demonstrated in two different classification tests.

1 Introduction

Microarrays represent a widely employed tool in molecular biology and genetics, which have produced an enormous amount of data to be processed to infer knowledge. Computational methodologies may be very useful in such analysis: among others, clear examples are tools aiding the microarray probe design, image processing-based techniques for the quantification of the spots, segmentation of spots/background, grid matching, noise suppression [5], methodologies for classification or clustering [22]. In this paper we focus on this last class of problems, and in particular on the samples classification task. In this context, many 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 e.g. [13,21].

In particular, very recently, a class of approaches have shown to be useful and discriminant in this context: the so called *topic* or *latent models* – the two most famous examples being the Probabilistic Latent Semantic Analysis (PLSA – [10])

[*] Corresponding author.

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 230–241, 2010.
© Springer-Verlag Berlin Heidelberg 2010

and the Latent Dirichlet Allocation (LDA – [3]). These powerful approaches have originally been introduced in the text analysis community for unsupervised topic discovery in a corpus of documents, in order to correlate the presence of a word to the particular topic discussed; the whole corpus of documents can then be described in terms of these topics. These techniques have also been largely applied in the computer vision community [4].

One of the main characteristics of this class of approaches is represented by their interpretability [7]: they can model a dataset in terms of hidden topics (or processes), which can reflect underlying and meaningful structures in the problem. This characteristic may be extremely useful in bioinformatics, where interpretability of methods and results is crucial. Topic models have already been applied in the context of expression microarray analysis: a tailored version of LDA (called Latent Process Decomposition – LPD), explicitly modelling expression levels, has been proposed in [19], with the aim of clustering expression microarray data; moreover, an application of topic models to biclustering has been recently proposed in [1].

A somehow unexplored scenario is represented by the application of such models in the classification context – a preliminary evaluation of standard topic models have been recently proposed in [2]. Even if supported by very promising results, a clear drawback is represented by the underlying basic assumption that each gene expression is independently generated given its corresponding latent topic.

In this paper a novel topic model is proposed, which we call BaLDA (Biologically-aware Latent Dirichlet Allocation), which starts from the Latent Process Decomposition [19], introduced in the context of clustering, and defines a new model able to take into account the given dependence between genes. This dependence is introduced in the graphical model through a variable, modeling a categorization of genes (namely a subdivision of genes in groups), which can be inferred by a priori knowledge on the genes of the analyzed problem. As a further refinement, a better modelling of the expression level is achieved by substituting the Gaussian pdf – present in the LPD – with a more descriptive Mixture of Gaussians.

We will show the usefulness of BaLDA in two classification experiments, assessing the impact of the different introduced modifications; a comparison with the LPD topic models and state of the art methods demonstrates the competitiveness of the proposed approach.

The rest of the paper is organized as follows: in Sec. 2 technical preliminaries about topic models are given. In Sec. 3 the model, together with learning/inference mechanism presented. An exhaustive experimental section is presented in Sec. 4, and, finally, in Sec. 5, we draw some conclusions.

2 Background

In this section the background concepts are reviewed. In particular, after introducing the general ideas underlying the family of topic models, we will present

Laten Dirichlet Allocation using the terminology and the notation of the document analysis context. Then we will briefly review how these models have been applied to the microarray scenario.

2.1 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 several topics (e.g. sport, finance, politics), which 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.

A variety of probabilistic topic models have been used to analyze the content of documents and the meaning of words. In the following section we will briefly present the LDA model, mainly to set up notations used in the remainder of the paper.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) was first introduced by Blei in [3]. In the LDA model, words are the only observable variables and they implicitly reflect a latent structure, i.e., the set of K topics used to generate the document. Generally speaking, given a set of documents, the latent topic structure lies in the set of words itself. In generating the document, for each word-position a topic is sampled and, conditioned from the topic, a word is selected. Each topic is chosen on the basis of the random variable θ that is sampled for convenience from a Dirichlet distribution $p(\theta|\alpha)$ where α is a hyperparameter. The topic z conditioned on θ and the word w conditioned on the topic and on β are sampled from multinomial distributions $p(z_n|\theta)$ and $p(w_n|z_n, \beta)$ respectively. β represents the word distribution over the topics. Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z_n, and a set of N words w_n that compose the document is given by

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \cdot \prod_{n=1}^{N} p(z_n|\theta) \cdot p(w_n|z_n, \beta) \tag{1}$$

where $p(z_n = i|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the probability of a document.

2.3 Topics Models in Bioinformatics

The representation provided by topic models has one clear advantage: each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms, see for example [6,2,19]. This may be really advantageous in the expression microarray context, since the final goal is to provide knowledge about biological systems, and discover possible hidden correlations. In particular there is a straightforward analogy between the pairs word-document and gene-sample: the expression level of a gene in a sample may be easily interpreted as the level of the presence of a word in a document (the higher the level the more present/expressed the word/gene is). In this sense, a particular topic model assumes that microarray data (represented as the gene-expression matrix) arises from a mixture of topics, whose number is fixed; changing the topic allows different subsets of genes to be prominent.

A possible problem which may arise is that expression microarray data is described with a matrix of real numbers, not as a non-negative integer matrix. This problem has been solved in [19] by modifying the standard LDA via the introduction of Gaussian distributions in place of word multinomial distributions β; this results in a novel and efficient probabilistic model called Latent Process Decomposition (LPD), where LDA topics are called "processes". The model has been successfully applied to clustering. Some modifications of the LPD model have been recently introduced: in particular, an optimized training version can be found in [23]; moreover, in [15], the LPD has been equipped with learned hyperpriors on the gaussian word-topic distributions. A method for maximizing lower bounds by re-estimating hyperparameters leaded to more accurate clustering results.

A somehow unexplored scenario is represented by the application of such models in the classification context; only very recently PLSA and LDA have been employed to classify expression microarray samples, with really promising results [2]. In particular, in [2], the original topic models [3,10] have not been changed; instead the gene expression matrix has been transformed, by a proper scaling and shifting, to a positive integer valued matrix, thus interpretable as a count matrix in the original LDA-PLSA formulation. Despite the method lacks biological motivations, it yielded very good classification results.

3 Biologically-aware Latent Dirichlet Allocation (BaLDA)

The main contribution of this paper is the definition a novel topic model for the analysis of expression microarray data, which directly improves the one provided in [19]. This novel topic model has two clear advantages with respect to the Latent Process Decomposition (LPD), detailed in the following.

The first (and most important) advantage starts from the observation that the major drawback of the PLSA, LDA and LPD models is the assumption that each gene expression is independently generated given its corresponding latent topic. While such representation provides an efficient computational method, it lacks the power to describe the coherent expression of different genes in a subset of samples, this aspect being widely known in the biology. In the proposed approach we include a mechanism in the graphical model that permits to include a priori knowledge on the relation between genes. This a priori information is expressed in terms of a gene categorization, namely a subdivision of the genes in groups of related genes based on external information, like known co-regulation,

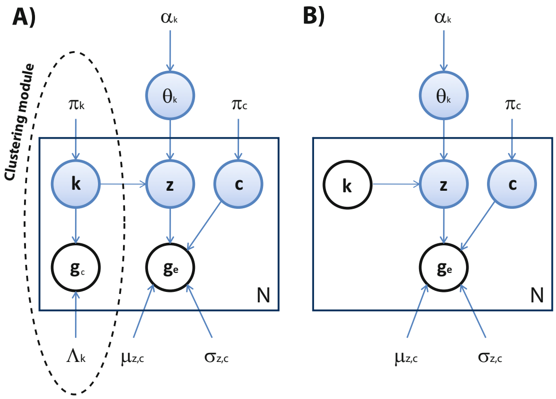

Fig. 1. A) Biologically-aware Latent Dirichlet Allocation Bayesian Network. Shaded/ Unshaded nodes are visible/hidden variables ($\mathbf{v}/\mathbf{h}$). The model parameters Ω are represented with a letter outside a node. B) A second version of the Biologically-aware Latent Dirichlet Allocation. The clustering result is fed into the model by means of the visible variable k.

spatial proximity or similarity of nucleotidic sequences to name a few. This categorization (i.e. clustering), which may be directly fed to the model, can be computed beforehand or can be simultaneously estimated while estimating the topic model.

The former option result in a straightforward modification of LDA; we add a visible variable k that influences the hidden topic variable z (see Fig. 1B). More interesting is the latter option, which permits LDA to deal with the uncertainty associated to the clustering. In this case (see Fig. 1A), the variable k is hidden, and depends on the visible variable g_c which represents the external information. These variables are modelled through a set of parameters which are learned simultaneously with the other parameters.

The second novelty of the proposed approach is related to the modelling of the word/topic distribution: in the original Latent Dirichlet Allocation, a word is generated by a multimodal distribution β, where $\beta_{w,z}$ represents the probability of finding the word w when the document is "speaking" about the topic z. In the LPD [19] the word-topic probability is modeled by a single gaussian, thus reflecting the continuous nature of the expression level, which is not captured with the original discrete formulation. Nevertheless, the monomodal nature of the Gaussian may not properly capture the possibly multimodal behavior of the gene-topic distribution: in particular, within a gene, a topic can be assigned to a single expression level. This limitation is removed in the proposed model, where the single Gaussian is replaced by a mixture of C Gaussians; which for large C, goes towards the multimodal spirit of the original multinomial β, still maintaining the appealing characteristic of modelling continuous expression levels.

3.1 BaLDA

The Bayesian network of Biologically-aware Latent Dirichlet Allocation (BaLDA) is depicted in Fig.1. The model is characterized by two observations, $g_{\mathbf{c}}$ and $g_{\mathbf{e}}$ (visible variables $\mathbf{v}$) which respectively govern the clustering and the topic submodules.

The variable k clusters the N genes in K components, while the parameters Λ_k represent the parameters of the particular probability density function chosen. For microarray expression are often used gaussians, t-distributions or factor analyzers [16]. We used gaussian clustering, so $\Lambda_k = \{\mu_k, \sigma_k\}$

$$p(g_{\mathbf{c},n}|k,\Lambda) = p(g_{\mathbf{c},n}|\Lambda_k) = \frac{1}{\sqrt{(2\pi)}\sigma} \cdot e^{\left(\frac{(g_{\mathbf{c},n}-\mu_k)^2}{-2\sigma_k^2}\right)} \tag{2}$$

The parameter π_k is a multinomial distribution that represents the prior on the cluster assignment.

Each n-th gene expression $g_{\mathbf{e},n}$ is assigned a topic $z_n = \{1 \ldots Z\}$ evaluating the gene-topic distribution and using a topic prior θ. We have that

$$p(g_{\mathbf{e},n}|z,\mu,\sigma) = \sum_{[c]} p(g_{\mathbf{e},n}|z,c,\mu,\sigma) = \sum_{[c]} \pi_{z,c,n} \cdot \frac{1}{\sqrt{(2\pi)}\sigma} \cdot e^{\left(\frac{(g_{\mathbf{e},n}-\mu_{z,c,n})^2}{-2\sigma_{z,c,n}^2}\right)} \tag{3}$$

where is now visible the mixture of Gaussians palette we introduced. With $[c]$ we indicate the values the variable c can assume. The prior on such topic assignment depends on the co-regulated genes (see the link $k \rightarrow z$ in the Bayesian network).

$$p(z = a|\theta, k) = \theta_{k,a} \tag{4}$$

where θ_k are multinomial distributions that represent the topic proportions used to generate each sample. Each distribution θ_k is governed by a Dirichlet prior $p(\theta_k|\alpha_k)$, where α is hyperparameter that represent the strength of a topic within a dataset.

$$p(\{\theta_k\}|\{\alpha_k\}) = \prod_{[k]} p(\theta_k|\alpha_k) = \prod_{[k]} \Big(\frac{1}{\mathcal{Z}(\alpha)} \prod_{z} \theta_{k,z}^{\alpha_k - 1} \Big) \tag{5}$$

Again the products are taken over the values of k and z and $\mathcal{Z}(\alpha)$ is Dirichlet distribution normalization constant.

At this point we can write the joint probability which describes the generative model as

$$p(g_{\mathbf{c}}, g_{\mathbf{e}}, c, k, z, \theta|\alpha, \mu, \sigma, \Lambda, \pi_c, \pi_k) = p(c|\pi_c) \cdot p(k|\pi_k) \cdot p(\theta|\alpha)$$
$$\prod_n \Big(p(g_{\mathbf{c},n}|k,\Lambda) \cdot p(g_{\mathbf{e},n}|c,z,\mu,\sigma) \cdot p(z_n|\theta) \Big)$$

where each conditional distribution has already been parameterized.

3.2 Inference and Learning

Under the model so far described, each t-th observation g^t is characterized by four hidden variables $\mathbf{h}^t = \{k^t, c^t, z^t, \theta_k^t\}$ which in turn are governed by the following parameters $\Omega = \{\Lambda_k, \pi_k, \mu_c, \sigma_c, \pi_c, \alpha\}$.

As in LDA, exact inference is intractable: we approach it using the variational inference [12]. We introduce a tunable distribution $q(\mathbf{h})$ over the hidden variables which defines the free energy $\mathcal{F}$

$$\mathcal{F} = \sum_t \Big(\sum_{\mathbf{h}} q(\mathbf{h}^t) \log \frac{q(\mathbf{h}^t)}{p(\mathbf{g}^t, \mathbf{h}^t | \Omega)} \Big) \tag{6}$$

We used the following form for the approximate posterior distribution, $q(\mathbf{h}^t) = q(\theta^t) \cdot \prod_n q(z_n^t, c_n^t) \cdot q(k_n^t)$ with $q(\theta_k^t)$ being a Dirac function centered at the optimal vectors $\hat{\theta}^t$. After plugging the approximate posterior and the joint distribution in the free energy formulation, we can iteratively decrease $\mathcal{F}$ with the Expectation-Maximization (EM) algorithm. The EM algorithm alternates in minimizing the free energy with respect to $q(\mathbf{h})$ (*E-Step*) and with respect to the model parameters Ω (*M-Step*). When updating q, the only constraint is that $\sum_{h_i} q(h^t) = 1$ for each hidden variable h and for each sample t. The update rules are simply obtained by setting the derivatives of $\mathcal{F}$ equal to zero and this reduces to the following formulas:

$$q(z_n^t = a, c_n^t = b) \propto \pi_b \cdot \mathcal{N}(g_{\mathbf{e},n}; \mu_{a,b,n}, \sigma_{a,b,n}) \cdot e^{\left(\sum_{[k_n]} q(k_n^t) \cdot \left(\Psi(\hat{\theta}_{b,a}) - \Psi(\sum_{[k]} \hat{\theta}_{k,b})\right)\right)} \tag{7}$$

where Ψ is the derivative of the logΓ function, computable via Taylor approximation (for further details see [3]), and $\mathcal{N}$ is the normal probability function (see Eq.2). The remaining updates of the E-step are

$$\hat{\theta}_{b,a}^t \propto \alpha_{b,a} + \sum_n q(k_n^t = b) \cdot q(z_n^t = a) \tag{8}$$

$$q(k_n^t = k) \propto \pi_k \cdot \mathcal{N}(g_{\mathbf{t},n}; \mu_k, \sigma_k) \tag{9}$$

In the M-step the collected posterior distributions q are used to compute an estimate $\hat{\Omega}$ of the model parameters

$$\mu_{n,c,z} = \frac{\sum_t q(z_n = z) \cdot q(c_n^t = c) \cdot g_{\mathbf{e},n}^t}{\sum_t q(z_n = z) \cdot q(c_n^t = c)} \tag{10}$$

$$\sigma_{n,c,z}^2 = \frac{\sum_t q(z_n = z) \cdot q(c_n^t = c) \cdot (g_{\mathbf{e},n}^t - \mu_{n,c,z})^2}{\sum_t q(z_n = z) \cdot q(c_n^t = c)} \tag{11}$$

$$\pi_{c,z,n} = \sum_t q(c_n = c) \cdot q(z_n = z) \tag{12}$$

The appropriate update on topic proportions' priors α_k can be obtained using a gradient descend

$$\{\hat{\alpha}_{k,a}\} = \arg\max \sum_t (\alpha_{k,a} - 1) \log \theta_{k,a} \tag{13}$$

subject to the appropriate normalization constraint.

We omit the update formulas for μ_k, σ_k^2 and π_k which can be computed in a very similar fashion.

3.3 Expression Microarray Samples Classification

In general, topic models have been originally introduced for clustering sets of documents: given the dataset, models are trained and analyzed in order to find clusters. Nevertheless, recently, they have been also successfully employed in the classification scenario – see for example [4,2]. The main idea is to employ a hybrid generative-discriminative approach [11], which exploits the generative model to extract a set of features to be classified with a discriminative classifier. More in detail, the training phase is carried out by first learning the models on the training set. Then a set of features is extracted from each sample; the transformed training set is then used to train a classifier. In the testing phase, the same feature extraction process is applied to the test sample, resulting in a feature vector to be classified using the trained classifier. In our work we employed the scheme proposed in [4,2], i.e. we employ the mixture of topics θ^t as sample descriptor. This have been demonstrated to be really discriminant [4,2]. Another benefit of this representation is that we are reducing the dimensionality from the number of genes N to the number of topics K, with $K \ll N$ – thus providing a compact and more interpretable representation. Finally, we are describing samples with a multinomial distribution whose characteristics will be exploited by the particular chosen classifiers.

4 Experiments

The proposed classification scheme has been evaluated using two different datasets, both related to tumors. The first derives from a study of prostate cancer by Dhanasekaran et.al [20], and consists of 54 samples with 9984 features. Such samples are subdivided in different classes: 14 samples are labelled as benign prostatic hyperplasia (labelled BPH), 3 as normal adjacent prostate (NAP), 1 as normal adjacent tumor (NAT), 14 as localized prostate cancer (PCA), 1 prostatitis (PRO) and 20 as metastatic tumors (MET). The 6 classes can be divided in three macro-classes: non-cancer (BPH,NAP,PRO), cancer (NAT,PCA), metastatic tumor (MET). This dataset has been also employed by the authors of [19] in their study for LPD. The second dataset we employed contains the expressions of 90 brain tissues used to study central nervous system embryonal tumor [18]. Each sample is characterized by 5920 features. The 90 samples include 60 with medulloblastomas, 10 with malignant gliomas, 5 with AT/RTs, 5 with renal/extrarenal rhabdoid tumors, 6 with supratentorial PNETs, and 4 normal cerebellum (5 classes in total). As in many expression microarray analysis, a beneficial effect may be obtained by selecting a sub group of genes, in order to limit the dimensionality of the problem and to reduce the possible redundancy present in the dataset. Here, as in [19], we decided to perform the experiments filtering the genes by variance and keeping only the top 500 genes.

In all the experiments we set $g_{\mathbf{c}} = g_{\mathbf{e}}$, namely we clustered the genes by looking at their expression levels in all the samples. This choice of course does not exploit the full potentiality of the method, but it permits to already obtain promising results (see tables below). Currently we are planning to perform an experiment by fully exploiting the potentialities of the model, considering different information (like spatial proximity or sequence similarity). In all the experiments, Z, K, and C, representing the number of topics, the number of clusters and the number of components in the mixture of Gaussians, respectively, are set in the following way: Z was found by applying the hold out log likelihood procedure described in [19], K has been automatically determined using Affinity Propagation [9] and $C = 3$ has been set after several tests.

In order to capture the different contributions of the two innovations of the model, we also tested the model with *i)* the clustering module but with only one Gaussian per gene (C=1), *ii)* the model enriched by the mixture of Gaussians gene-topic distribution, without the clustering information (K=1). We will refer to these two versions as BaLDA v1 and BaLDA v2 respectively.

The extracted features have been classified using Support Vector Machines employing a variety of kernels. Beside the standard linear kernel (LI), the probabilistic nature of the extracted features has been exploited by the use of different kernels on measures – also called information theoretic kernels [14], which provide similarity between probabilistic distributions; we employed some recent kernels, like the Kullbach-Leibler (KL), the Jensen-Shannon (JS) and the Jeffries kernels (JE). Finally we report also results with the K- Nearest Neighbor rule, using an approach similar to [2].

The proposed model has been compared with [19,2]. Even if [19] was designed for clustering data, it can be straightforwardly adapted to the classification scenario, following exactly the same hybrid scheme we employed. In order to have a fair comparison, we used the authors' implementation. Moreover, for a given choice of (K, Z) in BaLDA, we trained two LPD models: one with the same number of topics $Z_{LPD} = Z$, and one with the same complexity $Z_{LPD} = K \cdot Z$; this permits to give to the LPD the same number of processes that we have in our model. It is important to notice that the optimal Z for LPD, found by applying to the hold out log likelihood procedure, has been used also for BaLDA. In fact it is not obvious that the optimal Z_{LPD} will be the optimal for BaLDA as well. Classification errors have been computed using 10-fold cross validation (with 40 repetitions). In order to augment the statistical significance of the results, the generative models have been trained 4 times and results averaged.

Results, for both datasets, are reported in Table 1 and 2, respectively. From the tables it is evident the improvement obtained with the BaLDA models. In particular, in all the provided experiments the full model is performing better than the original LPD model (except in one case), with very remarkable improvements in the first dataset, also employed in the original paper of [19]. Moreover, by comparing the results of BaLDA v1 and BaLDA v2, we can observe that the improvement introduced by clustering the genes is more relevant than the other; however the combination of the two eventually yielded the best

Table 1. Results obtained from Prostate Cancer Dataset. See the text for the kernel abbreviations. We tested [2] also using the information theoretic kernels reporting the accuracies for the best Z.

	Z	K	C	LI	KL	JS	JE	KNN
LPD [19]	3	n.a.	n.a.	65.41	66.04	68.55	68.55	77.70
LPD [19]	12	n.a.	n.a.	86.16	82.39	85.53	85.53	82.22
[2]	3	n.a.	n.a.	82.38	83.64	78.60	84.90	77.89
BaLDA v1	3	4	1	86.80	88.68	88.05	89.94	88.17
BaLDA v2	3	1	3	77.98	76.73	76.73	75.47	76.67
BaLDA	3	4	3	89.94	89.31	91.20	**91.20**	85.24

Table 2. Results obtained from Brain Tumor Dataset. On the bottom, we reported the best accuracies of three other state of the art methods.

	Z	K	C	LI	KL	JS	JE	KNN
LPD [19]	15	n.a.	n.a.	83.33	81.48	81.85	84.07	78.56
LPD [19]	90	n.a.	n.a.	66.67	66.67	66.67	66.67	82.11
BaLDA v1	15	6	1	85.56	85.56	88.15	88.52	82.74
BaLDA v2	15	1	3	76.67	84.08	76.67	80.37	76.48
BaLDA	15	6	3	85.19	85.19	87.87	**88.89**	81.15

Comparison with the state of the art

Method	Acc.	Method	Acc.	Method	Acc.
[17]	86.50	[8]	86.20	[2]	84.1

result. Considering the classifiers, it is not clear which is the best combination of kernels and classifiers – this depending on the given dataset and on the given generative model. As a general comment, it can be said that information theoretic kernels are working better than the linear one, so confirming the intuition that exploiting the probabilistic nature of the features may be useful.

A final comment regards the interpretability of the method. Figure 2 describes topic proportions of the different models. We can observe that the topics can capture the different classes of the problem (with our model producing a qualitative better result – for more comments see the caption of the figure). This appealing interpretability of the topic models has been recently exploited in a biclustering scenario (see [1]).

5 Conclusions

In this paper we proposed a novel topic model, which enriches and extends the Latent Dirichlet Allocation (LDA) model by integrating genes' dependencies, encoded in a categorization of genes which better models the gene-topic distribution, leading to better classification of samples. The proposed model, called

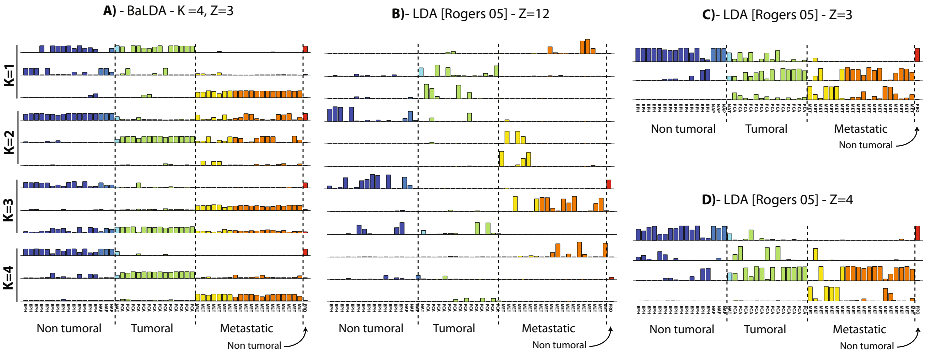

Fig. 2. Topic proportions θ of the prostate cancer dataset. We depict each of the classes with different colors. A) By clustering the genes BaLDA is able to use different topics to describe the 3 macroclasses; for example for the genes of the fourth cluster (K=4), the first topic describes the non-tumoral samples, the second topic the tumoral samples and the third the metastatic tumors. Again other clusters seem to highlight one of the three classes (the third cluster – K=3 – highlights metastatic using topic 2, etc). B) Comparison with [19] using a model with the same complexity. C) Comparison with [19] using the same number of topics. D) Comparison with [19] using the optimal topic number.

BaLDA has used to derive a highly informative and discriminant representation for microarray experiments. An experimental evaluation of the proposed methodologies on standard datasets confirms the effectiveness of the proposed techniques, also in comparison with other classification methodologies. Future works will focus on the biological interpretation of the results; it is evident that the interpretable topic representation of the expression matrix can be exploited to highlight genes strictly involved in the biological problem of interest, e.g. cancer or tumoral processes [1].

Acknowledgements

We acknowledge financial support from the FET programme within the EU FP7, under the SIMBAD project (contract 213250).

References

1. Bicego, M., Lovato, P., Ferrarini, A., Delledonne, M.: Biclustering of expression microarray data with topic models. In: Proc. Int. Conf. on Pattern Recognition (2010)
2. Bicego, M., Lovato, P., Oliboni, B., Perina, A.: Expression microarray classification using topic models. In: ACM SAC - Bioinformatics and Computational Biology track (2010)
3. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. of Machine Learning Research 3, 993–1022 (2003)

4. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via PLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
5. Brändle, N., Bischof, H., Lapp, H.: Robust DNA microarray image analysis. Machine Vision and Applications 15, 11–28 (2003)
6. Castellani, U., Perina, A., Murino, V., Bellani, M., Brambilla, P.: Brain morphometry by probabilistic latent semantic analysis. In: MICCAI (2010)
7. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.: Reading the tea leaves: how humans interpret topic models. In: NIPS (2009)
8. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(1), 3 (2006)
9. Frey, B., Dueck, D.: Clustering by passing messages between data points. Science 315, 972–976 (2007)
10. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42(1-2), 177–196 (2001)
11. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: NIPS, pp. 487–493 (1999)
12. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. Machine Learning 37(2), 183–233 (1999)
13. Lee, J., Lee, J., Park, M., Song, S.: An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48(4), 869–885 (2005)
14. Martins, A., Smith, N., Xing, E., Aguiar, P., Figueiredo, M.: Nonextensive information theoretic kernels on measures. J. of Machine Learning Research 10, 935–975 (2009)
15. Masada, T., Hamada, T., Shibata, Y., Oguri, K.: Bayesian multi-topic microarray analysis with hyperparameter reestimation. In: Proc. Int. Conf. on Advanced Data Mining and Applications (2009)
16. McLachlan, G., Bean, R., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. BMC Bioinformatics 18(3), 413–422 (2002)
17. Osareh, A., Shadgar, B.: Classification and diagnostic prediction of cancers using gene microarray data analysis. J. of Applied Sciences 9(3) (2009)
18. Pomeroy, S., Tamayo, P., et al.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415(6870), 436–442 (2002)
19. Rogers, S., Girolami, M., Campbell, C., Breitling, R.: The latent process decomposition of cdna microarray data sets. IEEE/ACM Trans. on Comp. Biology and Bioinformatics 2(2), 143–156 (2005)
20. Dhanasekaran, S., Barrette, T., et al.: Delineation of prognostic biomarkers in prostate cancer. Nature 23 412(6849), 822–826 (2001)
21. Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., Levy, S.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21(5), 631–643 (2005)
22. Valafar, F.: Pattern recognition techniques in microarray data analysis: A survey. Annals of the New York Academy of Sciences 980, 41–64 (2002)
23. Ying, Y., Li, P., Campbell, C.: A marginalized variational bayesian approach to the analysis of array data. BMC Proceedings 2(suppl. 4), S7 (2008)