Investigating Topic Models' Capabilities in Expression Microarray Data Classification

Manuele Bicego, Pietro Lovato,
Alessandro Perina, Marianna Fasoli,
Massimo Delledonne,
Mario Pezzotti,
Annalisa Polverari,
and Vittorio Murino

Abstract—In recent years a particular class of probabilistic graphical models—called topic models—has proven to represent an useful and interpretable tool for understanding and mining microarray data. In this context, such models have been almost only applied in the clustering scenario, whereas the classification task has been disregarded by researchers. In this paper, we thoroughly investigate the use of topic models for classification of microarray data, starting from ideas proposed in other fields (e.g., computer vision). A classification scheme is proposed, based on highly interpretable features extracted from topic models, resulting in a hybrid generative-discriminative approach; an extensive experimental evaluation, involving 10 different literature benchmarks, confirms the suitability of the topic models for classifying expression microarray data.

Index Terms—Expression microarray, topic models, hybrid generative discriminative approaches

1 Introduction

Microarrays represent a widely employed tool in molecular biology and genetics, resulting in an enormous amount of data to be processed to infer knowledge. Computational methodologies may be very useful in such analysis: among others, a relevant class is represented by methodologies for classification or clustering [1], [2]. In this context, in recent years some promising techniques [3], [4], [5], [6] were based on a particular class of probabilistic approaches, called topic models. Such models have been imported from the text analysis realm as workhorses in several scientific fields [7], [8]. Their wide usage is motivated by their simplicity and expressiveness in dealing with very large data sets [9], thus being a convenient tool for the microarray data analysis problem. In this context, topic models have been mainly used for clustering: for example, a specific topic model, called Latent Process Decomposition (LPD), has been proposed in [3] for clustering genes (some extensions have been proposed by Ying et al. [4] and Masada et al. [5]). An approach based on the probabilistic Latent Semantic Analysis (pLSA—[10]) was proposed in [11], aimed at clustering gene expressions and other information in order to find regulatory modules. A further example can be found in [6], where the biclustering issue was faced with the pLSA. On the other side, in the specific context of expression microarray data analysis, the classification issue has been almost completely disregarded by researchers, thus not exploiting all the potentialities of such models. In fact, even if topic models have been introduced for clustering purposes, there exist some variants able to deal with the classification. For example, DiscLDA [12] or Supervised LDA [13] are able to explicitly take into account the label information while building the model. Another successful line of research is represented by the so-called generative embedding schemes (or score spaces), where topic models are trained in a standard way and exploited to map the objects to be classified into a feature space, where a discriminative classifier can be used. This belongs to the more general class of hybrid generative-discriminative classification approaches [14], a recent class of techniques aimed at taking advantage of the best of the generative and the discriminative paradigms—the former based on probabilistic class models and a priori class probabilities, learned from training data and combined via Bayes law to yield posterior probabilities, the latter aimed at learning class boundaries or posterior class probabilities directly from data, without relying on generative class models [15]. Within the context of generative embeddings, it has been largely shown in other fields that powerful and discriminative descriptors may be extracted from topic models [7], [8], [16]; such analysis is completely missing in the microarray context, where the potentialities of the topic models have not completely been exploited.

In this paper we fill this gap, by investigating the capabilities of topic models-derived feature vectors for the classification of microarray data, resulting in a hybrid generative discriminative classification scheme. The proposed approach has been tested on several different data sets; obtained results, compared with the state of the art and with other supervised variants, confirm the suitability of such models for the classification of expression microarray data. Some considerations on the interpretability of the obtained feature descriptors have been also provided, with the use of a real data set involving different species of grape plants.

The remainder of the paper is organized as follows: in Section 2, the theory under the topic models is reviewed, whereas the proposed approach is presented in Section 3, detailing how discriminant features may be extracted from topic models; in that section an example highlighting their interpretability is also proposed. The experimental evaluation is presented in Section 4 and discussed in Section 5; finally in Section 6 conclusions are drawn and future perspectives are envisaged.

2 Topic Models and Microarray

Topic models have been originally introduced in the text analysis community, in order to describe and model a set of documents. The basic idea underlying these methods is that each document may be characterized by the presence of one or more topics (e.g., sport, finance, politics), which induces 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.

The application of topic models in the expression microarray scenario starts from the analogy that can be set between the pair word document and the pair gene sample: actually it is reasonable to intend the samples as documents and the genes as word occurrences. In fact, each sample is characterized by a vector of gene expressions: the expression level of a gene in a sample may be easily interpreted as the count of words in a document (the higher the level the more present the gene/word is in the sample/document). The representation of documents/samples and words/genes with 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. This may be really advantageous in the expression microarray context, since the final goal is to provide knowledge about biological systems, and to suggest possible hidden correlations.

- *M. Bicego and P. Lovato are with the Dipartimento di Informatica, Università degli Studi di Verona, Ca' Vignal 2, Strada Le Grazie 15, 37134 Verona, Italy. E-mail: {manuele.bicego, pietro.lovato}@univr.it.*
- *A. Perina is with Microsoft Research, One Microsoft Way, Redmond, WA 98052. E-mail: alperina@microsoft.com.*
- *M. Fasoli, M. Delledonne, M. Pezzotti, and A. Polverari are with the Dipartimento di Biotecnologie, Università degli Studi di Verona, Ca' Vignal 1, Strada Le Grazie 15, 37134 Verona, Italy. E-mail: {marianna.fasoli, massimo.delledonne, mario.pezzotti, annalisa.polverari}@univr.it.*
- *V. Murino is with Istituto Italiano di Tecnologia (IIT), Pattern Analysis & Computer Vision (PAVIS), Via Morego 30, 16163 Genova, Italy. E-mail: vittorio.murino@iit.it.*

Manuscript received 2 Dec. 2011; revised 22 Aug. 2012; accepted 31 Aug. 2012; published online 17 Sept. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-12-0319.
Digital Object Identifier no. 10.1109/TCBB.2012.121.

In this paper, we employ two topic models: the probabilistic Latent Semantic Analysis [10] and the Latent Process Decomposition [3], a variant of Latent Dirichlet Allocation (LDA—[17]) specifically designed for microarray.

Even if such models have been introduced in the text analysis community, here we reformulated their theory in order to deal with the microarray scenario, assuming the analogy *gene/words*, *sample/documents*, and *expression-level/word-counts*.

2.1 Probablistic Latent Semantic Analysis

In the original formulation of probabilistic Latent Semantic Analysis [10], the input is a set of documents, each one containing a set of words. The documents are summarized by an occurrence matrix, where each entry indicates the number of occurrences of a given word in a given document. In the same way, in the microarray scenario we can assume as an input a set of D samples, summarized by an expression matrix $n(g_n, d)$ which measures the expression level of the gene g_n in the sample d. Suppose that we have N different genes appearing in the sample set. In pLSA, the presence of a gene g_n in the sample d is mediated by a latent *topic* variable, $z \in Z = \{z_1, \dots, z_Z\}$, also called *aspect* class, i.e.,

$$p(g_n, d) = p(d) \cdot \sum_z p(g_n|z) \cdot p(z|d). \tag{1}$$

In practice, the topic z_k is a probabilistic co-occurrence of genes encoded by the distribution $\beta_{z_k}(g) = p(g|z_k)$, $g = \{g_1, \dots, g_N\}$. $p(z|d)$ (with $z = \{z_1, \dots, z_Z\}$) represents the proportion of the topics in the sample d; finally $p(d)$ accounts for varying expression levels of the genes.

The hidden distributions of the model, $p(g|z)$ and $p(z|d)$, are learned using Expectation-Maximization (EM), maximizing the model data log-likelihood $\mathcal{L}$:

$$\mathcal{L} = \sum_{n=1}^{N} \sum_d n(g_n, d) \cdot \log(p(g_n, d)). \tag{2}$$

The E-step computes the posterior over the topics, $p(z|g, d)$, and the M-step updates the hidden distributions.

Once the model has been learned one can estimate the topic proportion of an unseen sample. Usually, the learning algorithm is applied fixing the previously learned parameters $p(g|z)$ and estimating $p(z|d)$ for the sample in hand. For a deeper review of pLSA, see [10].

2.2 Latent Process Decomposition

Latent Process Decomposition [3] represents a topic model which has been specifically designed for the microarray scenario. This model starts from the Latent Dirichlet Allocation [17], which has been introduced, like pLSA, on the linguist concepts of words, documents and topics.

In particular, the differences between LPD and pLSA are twofold: from one hand, LPD inherits from LDA the introduction of a Dirichlet prior on the mixture of topics that defines a document (sample), permitting a true generative model for the whole corpus of documents (samples) [18]. On the other hand, in the LPD, the gene-topic probability is modeled by a single gaussian $\langle \mu_{g,z}, \sigma_{g,z} \rangle$, thus reflecting the continuous nature of the expression level, which is not captured with the discrete formulation or the original LDA/pLSA.

The parameters of the model are two: α, namely the parameters of the Dirichlet distribution from which the random variable θ is sampled—θs are the topic proportions that define a document, namely the $p(z|d)$ for pLSA; and β, namely the word distribution over the topics—i.e., $p(g|z)$, for pLSA. Given these two parameters, the joint distribution of a topic mixture θ, a set of N topics z_n, and a set of N genes g_n expressed in the sample is given by

$$P(\theta, \mathbf{z}, \mathbf{g}|\alpha, \beta) = p(\theta|\alpha) \cdot \prod_{n=1}^{N} p(g_n|z_n, \beta) \cdot p(z_n|\theta), \tag{3}$$

where $p(z_n = k|\theta) = \theta(k)$, i.e., a multinomial evaluated in k.

For a deeper review of LPD and LDA please refer to [3] and [17], respectively.

3 Microarray Data Classification with Topic Models

This section describes how Topic Models may be used for classification of expression microarray data. In particular, we will describe how descriptive feature vectors can be extracted from them, giving also some insights about their interpretability.

As introduced in the previous section, given the analogy between the pair word document and the pair gene sample, we can in general associate the expression matrix of a microarray experiment to the count matrix $<w_j, d>$ of topic models, to be explicitly or implicitly used to train the specific probabilistic model.

A small remark: it is worth noting that gene expression is subject to complex coregulation mechanisms, and there are aspects of this interdependence that cannot be captured with words co-occurrence. Nevertheless, we will show later that our methods may work properly even if disregarding this biological aspect.[1]

3.1 The Feature Vector and the Classification Scheme

As explained in the introduction, the main idea is to employ a staged hybrid generative-discriminative approach, which, in our case, is realized as follows:

1. **Generative model training**: Given the training set, the generative topic model is trained, as explained in the previous section. Different schemes may be adopted to fit the best model (or set of models) to the data, namely by learning one model per class, one per the whole data set or others. Here, we employ the basic one, namely training one single model for all classes.
2. **Generative embedding**: Within this step, all the objects are projected, through the learned model, to a vector space. In particular, for a given experiment d, the representation $\phi(d)$ in the generative embedding space is defined as the estimated topic posteriors distribution (the $p(z|d)$ for pLSA and the analogous posterior Dirchlet parameter γ [3] for LPD). The intuition is that every topic may be approximately associated with a biological process (or to a set of—[3], [6]), which involves some particular genes and is active in particular samples. Thus, the topic distribution characterizing a sample may indicate which and to which extent the different processes are active in such sample, thus representing a significative and possibly discriminant feature. Moreover, it is important to notice that this representation with the topic posteriors has been already successfully used in computer vision tasks [20], [7] as well as in the medical informatics domain [21].
3. **Discriminative classification**: in the resulting generative embedding space any discriminative vector-based classifier may be employed. In this fashion, according to the generative/discriminative classification paradigm, we use the information coming from the generative process as discriminative features of a discriminative classifier.

These descriptors are advantageous from different points of view:

[1] A preliminary version of a topic model considering this aspect has been recently introduced by some of the authors in [19].

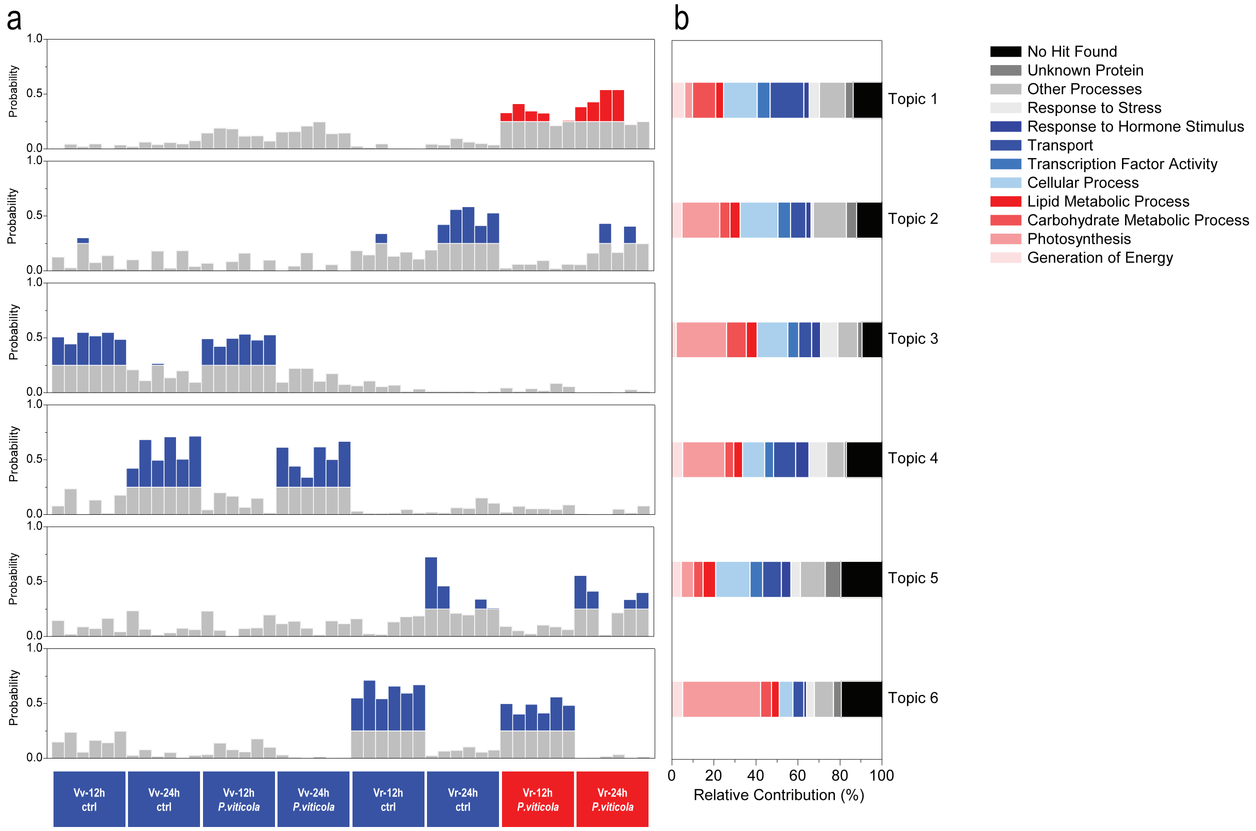

Fig. 1. PLSA analysis. (a) Bar representation of the $p(z|d)$ distribution for each of the six topics (each bar corresponds to one sample). The main classes are represented on the bottom of the figure. (b) Functional category distribution of topic specific genes.

1. in other contexts it has been shown that they are very descriptive for classification purposes [7], [20], [21] (this being confirmed by our experimental evaluation);
2. they provide a really interpretable representation of the microarray experiments, in terms of biological processes—see the next section for an example;
3. the dimensionality of the feature vector is reduced from the number of genes N to the number of topics K, with $K \ll N$—thus providing a more compact and easy-to-manage representation.
4. finally, such descriptors represent multinomial distributions,[2] which are suitable to be classified using kernels on probability measures (also called Information Theoretic Kernels)—which have been shown to be very effective in classification problems involving text, images, and other types of data (see [22] and the references therein); moreover, very recently, they have been shown to be very suitable for the hybrid generative-discriminative approach (see, for example, [23]).

3.2 The Interpretability of the Feature Vector

The extracted feature vectors are highly interpretable: in particular, $p(z|d)$ (θ_z^d) of PLSA (LPD) characterizes "how present" every topic is in a given sample; as shown in [3], [6], a topic may be easily associated with a biological process. Intuitively, a topic characterizes a subset of samples where the gene expressions are highly correlated. Therefore, $\phi(d)$ may be used to infer the different biological processes which are active over the different samples. It should be noted that also the probability of the genes given the topic—$p(g|z)$—may be very useful: actually it may be interpreted as the impact of the different genes in a particular biological process. Moreover, the probabilistic nature of these models permits to encode also the *level* of the impact, thus taking into account the well-known fact that not all biological processes are taking place in every sample.

To show these characteristics we applied the proposed scheme on a data set which included 48 samples (and 24,676 genes) of microarray expressions of two grapevine species, *V. vinifera* and *V. riparia*, both subjected to infection with *Plasmopara viticola*, a pathogen responsible for a destructive disease. It is known that *V. riparia* is resistant to the pathogen, while *V. vinifera* is more susceptible to infection, and the study focused in understanding molecular switches, signals, and effectors involved in resistance [24]. In the paper, they reported a microarray analysis of early transcriptional changes associated with *P. viticola* infection in both susceptible *Vitis vinifera* and resistant *Vitis riparia* plants (12 and 24 h post inoculation). The same experiments have been conducted with the plant treated with water, a neutral agent used as control. We choose this data set since it is very complex and structured; different classes can be highlighted: in particular, samples can be divided on the basis of the type of plant (*V. vinifera* or *V. riparia*), of the time point (after 12 or 24 h), or the pathogen/water treatment.

In the training phase, a pLSA model was trained 50 times and the best model (in a likelihood sense) was retained. Guided by the expertise of biologists, the number of topics has been set to 6 after several trials. Then, information have been extracted from the topic/document and word/topic distributions. In particular, in Fig. 1 we report on the left an intuitive bar plot of the probability $p(z|d)$ (different rows correspond to different topics z), while the figure on the right represents the functional categories of the most important genes (found by looking at $p(g|z)$—a counterpart for LPD may be somehow difficult to derive).

Studying the composition of the data set, we observed that it is rather accurately reflected by the $p(z|d)$ distribution (on the left of

[2] Actually, in the LPD case, we used the normalized posterior.

TABLE 1
Summary of the Employed Data Set

Dataset	G,S,C	Citation	Test Protocol
leuk2	11225,72,3	[25]	5-fold CV
leuk1	5327,72,3	[26]	10-fold CV
11tumors	12533,174,11	[27]	5-fold CV
colon	2000,62,2	[28]	LOO CV
brain1	5920,90,5	[29]	4-fold CV
brain2	10367,50,4	[30]	10-fold CV
lung	12600,203,5	[31]	5-fold CV
nci60	7129,60,9	[32]	10-fold CV
prostate	10509,102,2	[33]	LOO CV
9tumors	5726,60,9	[34]	10-fold CV

In particular, G represents the number of genes, S the number of samples, and C the number of classes.

the figure). Actually, every topic can reflect a different aspect of the data set. For example, some topics show groups of samples which are more correlated with the effects of treatment at the different time points rather than with a specific reaction to the pathogen in comparison with the control (water). This is evident in the third and fourth topics, which represent *V. vinifera* after 12 and 24 hour, respectively, the former without pathogen inoculation and the latter infected. The last topic captures the processes of *V. riparia* after 12 hour since the infiltration, in the first case with water, in the second with the pathogen.

From the specific disease resistance point of view, the analysis confirmed the tendency of a specific response in *V. riparia*. In fact, the first topic deals with samples related to infected *V. riparia*'s leaves at both time points (12 and 24 hours after infection). By looking at the genes which are most active in the first topic, biologists found that their distribution is particularly significant. In fact, important functional categories among the involved genes (listed on the right side of Fig. 1) are carbohydrate metabolism and transport, in contrast with a strong contribution of photosynthesis-related gene expression in other topics. As previously reported, the primary metabolic reprogramming underlies defense in biotrophic interactions in order to potentially supply both energy and precursors to implement a defense mode.

It is also worth noting that, within topic 1, the same trend of the last 12 experiments is visible on the classes of *V. vinifera* subjected to inoculation (samples 12-24). In fact, this means that an activation of some genes—possibly involved in the response to the pathogen—is undergoing, but the response is too weak, explaining the susceptibility of the plant to *P. viticola*.

Concluding, all these observations qualitatively confirm the capabilities of the proposed descriptors to encode different aspects of the data set. A quantitative evaluation is provided in the next section.

4 Experimental Evaluation

The suitability of the proposed classification scheme has been extensively tested on 10 different well-known data sets, briefly summarized in Table 1. The whole description of each data set may be found in the reported reference.

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 data set. Gene selection may be obtained using different methodologies, ranging from the simple variance filtering up to complicate statistics. Here, we employed the Minimum-Redundancy Maximum-Relevance feature selection approach [35], [36].[3] In order to have a fair comparison with the state of the art, for every data set we selected the best result in the literature (at least to the best of our knowledge)—they are reported in Table 1; we used then, in our experiments, the same number of genes used in that paper (when specified); if not specified, we retain 500 genes (as in the LPD paper [3]). For similar reasons, also the cross-validation protocols—again reported in Table 1—have been chosen by looking at the relative state-of-the-art papers.

In the learning phase, the pLSA and the LPD models have been built only on the training set. Since the training procedure can converge to local optima of the likelihood, the training has been repeated 20 times, starting from different random initializations, retaining the model with the highest data likelihood. The number of topic is a free parameter in topic models, and should be set in advance. Different automatic techniques have been proposed in the literature to set such a number, ranging from hold-out likelihood [3] to cross validation, from a priori knowledge to probabilistic model selection methods—e.g., the Bayesian Information Criterion (BIC—[46]). Here, we adopted a very simple scheme: starting from the observation that topic models were initially designed to discover and model groups of documents, we thought reasonable to fix the number of topics as proportional to the number of classes, namely the number of natural groups present in the data set (after few trials, we found that three times the number of classes was a reasonable choice). Despite the simplicity of this rule, obtained results were very satisfactory. An analysis of the performances of pLSA and LPD with respect to this parameter is presented in the discussion part. A final note on the training for the pLSA model: the expression matrix (real-valued numbers) cannot be used as it is as the count matrix $< w_j ; d_i >$ of topic models (positive and integer values); therefore a simple normalization step (shifting and scaling) has been applied to the matrix in order to have positive and integer values.

As almost always in hybrid generative discriminative schemes, the classification accuracies have been computed using Support Vector Machines in the resulting generative embedding space—the parameter C has been selected using Cross Validation on the training set. Here, more than using the standard linear kernel, we exploited the probabilistic nature of the feature vector by the use of different kernels on measures (also called information theoretic kernels [22]), which provide similarity between probabilistic distributions. It has been shown in other contexts (see, for example [23]) that such combination may be beneficial for some hybrid generative-discriminative methods. In particular, here we employ the standard Jensen-Shannon kernel (JS) and a novel kernel, recently introduced by Martins et al. [22], which is based on a nonextensive generalization of the classical Shannon information theory, and defined on (possibly unnormalized) probability measures (see [22] for all details): the Jensen-Tsallis (JT) kernel (the parameter q has been adjusted by cross validation on the training set).

In order to investigate the potentiality and the possible extendibility of the proposed approach, we extracted from the topic models a more complex feature, called Free Energy Score Spaces (FESS—[8]), which expresses how well each data point (i.e., microarray experiment) fits different parts of a trained generative model. It has been found that the FESS is highly informative for discriminative learning, yielding state-of-the-art results in several contexts [8], [47]. In our experiments, after extracting the FESS descriptors, we used the linear kernel with SVM.

Another interesting point of analysis is related to the different possible ways in which topic models can be exploited for classification. Alternatives to our staged scheme exist, as explained in the introduction: in particular, here we compared our approach to a simple Bayesian scheme—which trains one model per class and performs classification with the Bayes rule—, and to the supervised topic models approach [13]—which explicitly takes into account the labels in the training process.[4]

[3] http://www.mathworks.com/matlabcentral/fileexchange/14916.

[4] The code can be found in http://cran.r-project.org/web/packages/lda/.

TABLE 2
Classification Errors of the Proposed Approaches for Different Data Sets

Method	Leuk2	Leuk1	11Tumors	Colon	Brain1	Brain2	Lung	NCI60	Prostate	9Tumors
PLSA Lin	0.0267	0.0286	0.0900	0.0968	0.0982	0.1000	0.0541	0.4800	0.0686	0.3200
PLSA JS	0.0267	0.0143	0.0583	0.0968	0.0982	0.1200	0.0447	0.3067	0.0490	0.2533
PLSA JT	0.0267	0.0286	0.0534	0.1129	0.0967	0.2200	0.0397	0.3067	0.0490	0.2733
LPD Lin	0.0133	0.0125	0.0947	0.1935	0.1205	0.1800	0.0585	0.2700	0.0588	0.3433
LPD JS	0.0133	0.0393	0.0671	0.1935	0.1429	0.1800	0.0673	0.2533	0.0588	0.3600
LPD JT	0.0133	0.0143	0.0840	0.1774	0.1101	0.1800	0.0578	0.2867	0.0490	0.3767
FESS L2	0.0267	0.0278	0.0465	0.0833	0.0773	0.0778	0.0711	0.0963	0.0392	0.0516
FESS L3	0.0267	0.0417	0.0457	0.0833	0.0761	0.0778	0.0711	0.0963	0.0392	0.0556
Our Best	**0.0133**	**0.0125**	**0.0457**	0.0833	**0.0761**	**0.0778**	**0.0397**	**0.0963**	0.0392	**0.0516**
Bayesian Scheme	0.0297	0.0143	0.0847	0.1452	0.0863	0.2800	0.0542	0.3433	0.0882	0.2933
Supervised TM	0.0810	0.0833	0.5866	0.0806	0.3318	0.3200	0.1541	0.6733	0.0588	0.5900
State of the art (Ref)	0.0150 ([37])	0.0250 ([38])	0.0520 ([39])	**0.0650** ([40])	0.1350 ([41])	0.1500 ([38])	0.0620 ([42])	0.1170 ([43])	**0.0240** ([44])	0.2460 ([45])

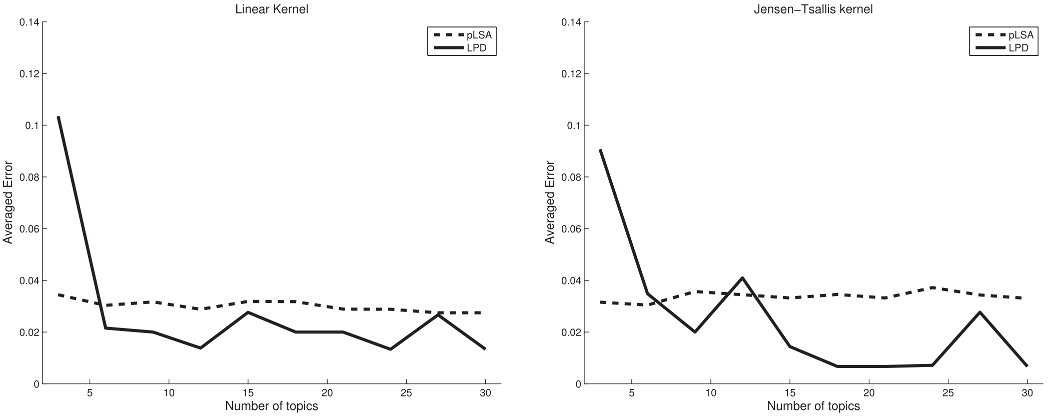

Fig. 2. Accuracies on the Leuk data set by varying the number of topics. Results are shown for pLSA and LPD with linear (left) and Jensen-Tsallis (right) kernels.

All the obtained results are reported in Table 2, together with state-of-the-art results (in bold the best result for every data set). “Lin,” “JS,” and “JT” stand for linear, Jensen-Shannon, and Jensen-Tsallis kernels, respectively. “FESS L2” and “FESS L3” are two variants of the FESS approach—see [8] for all details.

5 Discussion

As a general comment, from the table it can be argued that descriptors extracted from Topic Models are really effective for expression microarray classification. When compared with literature, we can observe that our results are in line with those results. Moreover, in three cases (Brain1, Brain2, and 9 Tumors), our best result is substantially better than the state of the art. It is important to notice, at this point, that we compared our results (obtained within a single framework) with results obtained with many different techniques on different data sets, each technique possibly tailored for the specific data set (which are very different in terms of composition and difficulty—see Table 1).

Some more specific observations can be drawn from the table: in particular, by looking at the behavior of the different kernels, we can notice that a beneficial effect is obtained when exploiting the probabilistic nature of the feature vector by using the information theoretic kernels. Concerning the two employed generative models, it seems that in average there is not such a big difference betwenn pLSA and LPD in terms of accuracy, with some data sets slightly preferring pLSA. A possible justification may be searched in the sensibility of LPD model to the choice of the number of topics. To investigate such behavior, we performed an exhaustive analysis on the Leuk2 data set, by varying such number from 3 to 30 (step 3). In Fig. 2, the error curves are displayed.

It seems evident from the plots that the accuracies for the pLSA do not vary too much while changing the number of topic, whereas the LPD is more sensible to such choice (when properly chosen, LPD outperforms pLSA). This is true both for linear and for the JT kernels.

The potentiality and the possible extendibility of the proposed approach are evident when looking at the results obtained with FESS. Actually, it turned out that when the topic proportion descriptor is somehow not enough to discriminate (see, for example, NCI60 and 9 tumors), the FESS signature permits to unravel the complexity of the problem, leading to excellent results (on the contrary, when the topic proportion feature vector works well, only a marginal improvement is got by using FESS).

Finally, by comparing the different ways of exploiting topic models for classification (our approach, the Bayesian scheme, and the supervised topic models method), it seems evident that in problems with few classes a supervised topic model is a good choice, leading to very good results. On the contrary, when the number of classes increases, the other two choices seem to be more appropriate. This is probably due to the fact that both approaches treat in a separate way the data and the labels (the Bayesian approach by splitting the training set, the hybrid approach by the two-staged procedure), whereas supervised topic models try to simultaneously consider both data and labels, which can be very complicated in problems with large number of classes. In general, our hybrid approach is better, confirming the fact, shown in other many different contexts, that this scheme is able to exploit the complementarity of the generative and the discriminative paradigms.

6 Conclusions

In this paper, we investigated the use of topic models for classification of expression microarray data. A classification scheme is proposed, based on highly interpretable features extracted from topic models, resulting in a hybrid generative-discriminative approach; an extensive experimental evaluation, involving 10 different literature benchmarks, confirmed the suitability of the topic models for classifying this kind of data. Finally, a qualitative analysis on grapevine plants expressions suggested the great expressiveness of the proposed approach.

REFERENCES

[1] J. Lee, J. Lee, M. Park, and S. Song, "An Extensive Comparison of Recent Classification Tools Applied to Microarray Data," *Computational Statistics & Data Analysis,* vol. 48, no. 4, pp. 869-885, 2005.

[2] M. de Souto, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep, "Clustering Cancer Gene Expression Data: A Comparative Study," *BMC Bioinformatics,* vol. 9, article 497, 2008.

[3] S. Rogers, M. Girolami, C. Campbell, and R. Breitling, "The Latent Process Decomposition of cDNA Microarray Data Sets," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 2, no. 2, pp. 143-156, Apr.-June 2005.

[4] Y. Ying, P. li, and C. Campbell, "A Marginalized Variational Bayesian Approach to the Analysis of Array Data," *BMC Proc.,* vol. 2, no. Suppl 4, article S7, 2008.

[5] T. Masada, T. Hamada, Y. Shibata, and K. Oguri, "Bayesian Multi-Topic Microarray Analysis with Hyperparameter Reestimation," *Proc. Int'l Conf. Advanced Data Mining and Applications,* 2009.

[6] M. Bicego, P. Lovato, A. Ferrarini, and M. Delledonne, "Biclustering of Expression Microarray Data with Topic Models," *Proc. Int'l Conf. Pattern Recognition,* pp. 2728-2731, 2010.

[7] A. Bosch, A. Zisserman, and X. Munoz, "Scene Classification via PLSA," *Proc. European Conf. Computer Vision,* vol. 4, pp. 517-530, 2006.

[8] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free Energy Score Space," *Proc. Neural Information Processing Systems,* 2009.

[9] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading the Tea Leaves: How Humans Interpret Topic Models," *Proc. Neural Information processing systems,* 2009.

[10] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning,* vol. 42, nos. 1/2, pp. 177-196, 2001.

[11] J. Joung, D. Shin, R. Seong, and B. Zhang, "Identification of Regulatory Modules by Co-Clustering Latent Variable Models: Stem Cell Differentiation," *Bioinformatics,* vol. 22, no. 16, pp. 2005-2011, 2006.

[12] S. Lacoste-Julien, F. Sha, and M. Jordan, "Disclda: Discriminative Learning for Dimensionality Reduction and Classification," *Proc. Information Processing Systems Conf.,* 2008.

[13] D. Blei and J. McAuliffe, "Supervised Topic Models," *Proc. Neural Information Processing Systems,* 2007.

[14] J. Lasserre, C. Bishop, and T. Minka, "Principled Hybrids of Generative and Discriminative Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[15] Y.D. Rubinstein and T. Hastie, "Discriminative vs Informative Learning," *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining,* pp. 49-53, 1997.

[16] J. Chappelier and E. Eckard, "Plsi: The True Fisher Kernel and Beyond," *Proc. European Conf. Machine Learning and Knowledge Discovery in Databases: Part I,* pp. 195-210, 2009.

[17] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[18] M. Girolami and A. Kabán, "On an Equivalence Between Plsi and Lda," *SIGIR '03: Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval,* pp. 433-434, 2003.

[19] A. Perina, P. Lovato, V. Murino, and M. Bicego, "Biologically-Aware Latent Dirichlet Allocation (Balda) for the Classification of Expression Microarray," *Proc. Int'l Conf. Pattern Recognition in Bioinformatics,* pp. 230-241, 2010.

[20] M. Cristani, A. Perina, U. Castellani, and V. Murino, "Geo-Located Image Analysis using Latent Representations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[21] U. Castellani, A. Perina, V. Murino, M. Bellani, and P. Brambilla, "Brain Morphometry by Probabilistic Latent Semantic Analysis," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention,* 2010.

[22] A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo, "Nonextensive Information Theoretic Kernels on Measures," *J. Machine Learning Research,* vol. 10, pp. 935-975, 2009.

[23] M. Bicego, A. Perina, V. Murino, A. Martins, P. Aguiar, and M. Figueiredo, "Combining free Energy Score Spaces with Information Theoretic Kernels: Application to Scene Classification," *Proc. IEEE Int'l Conf. Image Processing,* pp. 2661-2664, 2010.

[24] M. Polesani, L. Bortesi, A. Ferrarini, A. Zamboni, M. Fasoli, C. Zadra, A. Lovato, M. Pezzotti, M. Delledonne, and A. Polverari, "General and Species-Specific Transcriptional Responses to Downy Mildew Infection in a Susceptible (Vitis Vinifera) and a Resistant (v. Riparia) Grapevine Species," *BMC Genomics,* vol. 11, article 117, 2010.

[25] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den Boer, M. Minden, S. Sallan, E. Lander, T. Golub, and S. Korsmeyer, "MLL Translocations Specify a Distinct Gene Expression Profile that Distinguishes a Unique Leukemia," *Nature Genetics,* vol. 30, no. 1, pp. 41-47, 2002.

[26] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science,* vol. 286, no. 5439, pp. 531-537, Oct. 1999.

[27] A. Su, J. Welsh, L. Sapinoso, S. Kern, P. Dimitrov, H. Lapp, P. Schultz, S. Powell, C. Moskaluk, H.F. Frierson Jr., and G. Hampton, "Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures," *Cancer Research,* vol. 61, pp. 7388-7393, 2001.

[28] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA,* vol. 96, no. 12, pp. 6745-6750, 1999.

[29] S. Pomeroy, P. Tamayo, M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, J. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. Louis, J. Mesirov, E. Lander, and T. Golub, "Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression," *Nature,* vol. 415, pp. 436-42, 2002.

[30] C. Nutt, D. Mani, R. Betensky, P. Tamayo, J. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. McLaughlin, T. Batchelor, P. Black, A. von Deimling, S. Pomeroy, T. Golub, and D. Louis, "Gene Expression-Based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification," *Cancer Research,* vol. 63, no. 7, pp. 1602-1607, 2003.

[31] A. Bhattacherjee et al., "Classification of Human Lung Carcinomas by Mrna Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 13 790-13 795, 2001.

[32] D. Ross, U. Scherf, M. Eisen, C. Perou, P. Spellman, V. Iyer, S. Jeffrey, M. de Rijn, M. Waltham, A. Pergamenschikov, J. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Botstein, and P. Brown, "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics,* vol. 24, pp. 227-234, 2000.

[33] D. Singh et al., "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell,* vol. 98, pp. 203-209, 2002.

[34] J. Staunton, D. Slonim, H. Coller, P. Tamayo, M. Angelo, J. Park, S.U.J. Lee, W. Reinhold, J. Weinstein, J. Mesirov, E. Lander, and T. Golub, "Chemosensitivity Prediction by Transcriptional Profiling," *Proc. Nat'l Academy of Sciences USA,* vol. 98, no. 19, pp. 10787-10792, 2001.

[35] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Proc. IEEE CS Bioinformatics Conf.,* pp. 523-529, 2003.

[36] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[37] N. Yukinawa, S. Oba, K. Kato, and S. Ishii, "Optimal Aggregation of Binary Classifiers for Multiclass Cancer Diagnosis Using Gene Expression Profiles," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 6, no. 2, pp. 333-343, Apr.-June 2009.

[38] A. Statnikov, C. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis," *Bioinformatics,* vol. 21, no. 5, pp. 631-643, 2005.

[39] J. del Coz, J. Diez, and A. Bahamonde, "Learning Nondeterministic Classifiers," *J. Machine Learning Research,* vol. 10, pp. 2273-2293, 2009.

[40] H. Liu, L. Liu, and H. Zhang, "Ensemble Gene Selection by Grouping for Microarray Data Classification," *J. Biomedical Informatics,* vol. 43, no. 1, pp. 81-87, 2010.

[41] A. Osareh and B. Shadgar, "Classification and Diagnostic Prediction of Cancers Using Gene Microarray Data Analysis," *J. Applied Sciences,* vol. 9, no. 3, pp. 459-468, 2009.

[42] P. Chen, S. Huang, W. Chen, and C. Hsiao, "A New Regularized Least Squares Support Vector Regression for Gene Selection," *BMC Bioinformatics,* vol. 10, article 44, 2009.

[43] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, "Feature Selection for Gene Expression Using Model-Based Entropy," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 7, no. 1, pp. 25-36, Jan.-Mar. 2010.

[44] X. Wang and O. Gotoh, "A Robust Gene Selection Method for Microarray-Based Cancer Classification," *Cancer Informatics,* vol. 9, pp. 15-30, 2010.

[45] X. Hang, "Cancer Classification by Sparse Representation Using Microarray Gene Expression Data," *Proc. Bioinformatics and Biomedicine Workshops (BIBMW),* pp. 174-177, 2008.

[46] G. Schwarz, "Estimating the Dimension of a Model," *The Ann. of Statistics,* vol. 6, no. 2, pp. 461-464, 1978.

[47] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "A Hybrid Generative/Discriminative Classification Framework Based on Free-Energy Terms," *Proc. IEEE 12th Int'l Conf. Computer Vision,* 2009.